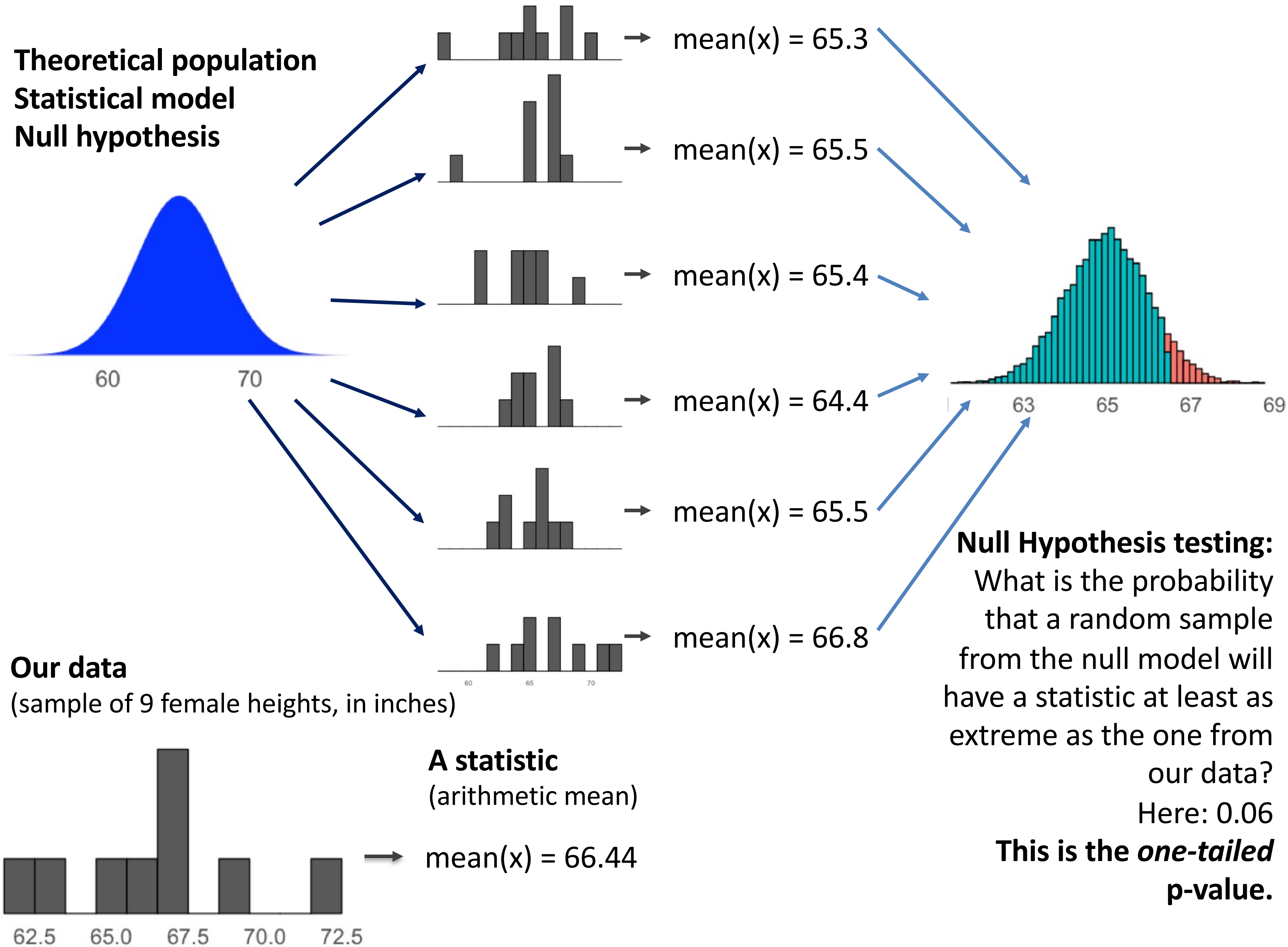
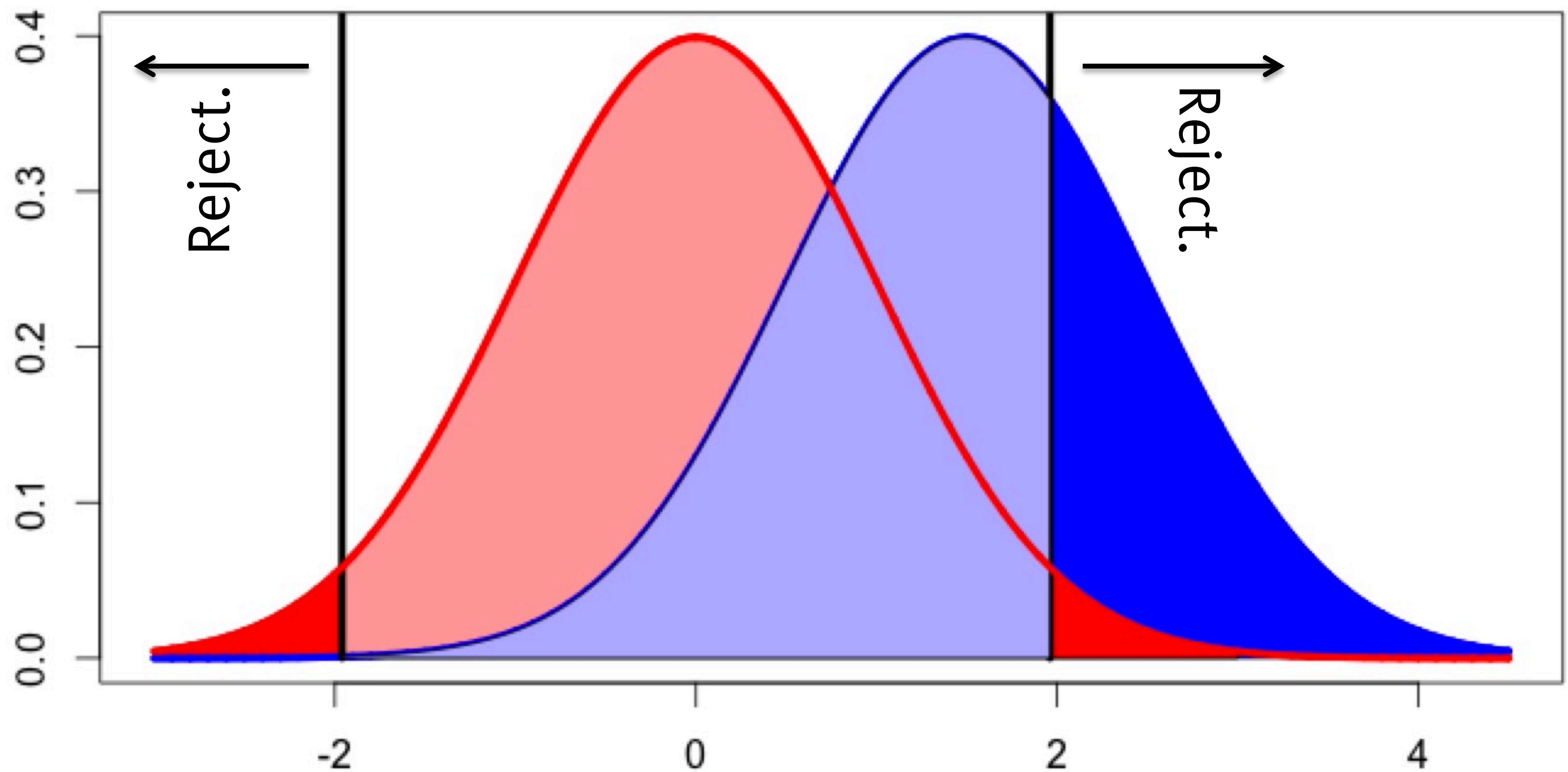
201ab Quantitative methods
L.07: common tests
t-test, chi^2, binomial



Theoretical population
Statistical model
Null hypothesis
60
70
mean(x) = 65.3
mean(x) = 65.5
mean(x) = 65.4
mean(x) = 64.4
mean(x) = 65.5
mean(x) = 66.8
60
65
70
63
65
67
69
Null Hypothesis testing:
What is the probability that a random sample from the null model will have a statistic at least as extreme as the one from our data?
Here: 0.06
This is the one-tailed p-value.
Our data
(sample of 9 female heights, in inches)
62.5
65.0
67.5
70.0
72.5
A statistic
(arithmetic mean)
mean(x) = 66.44

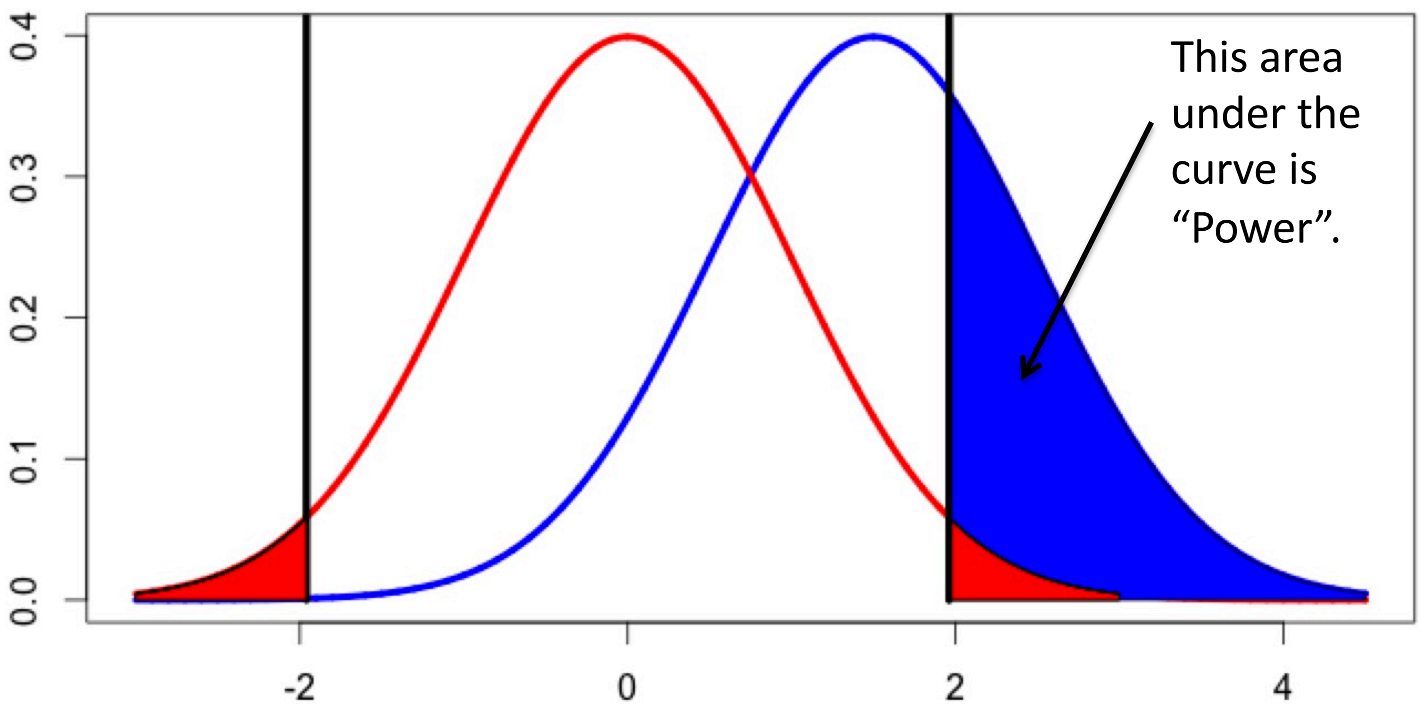
Errors in NHST



	H_0 false	H_0 true
Reject H_0	Correct rejection of null (Pr = 1-β 'power')	Type I error (Pr = α)
Fail to reject H_0	Type II error (Pr = β)	Correct failure to reject null (Pr = 1-α)

Power P(significant | not null)

- The conditional probability of rejecting the null hypothesis when the data actually came from the ‘alternate’ hypothesis distribution.
- To calculate this, we need to know what the ‘true effect’ distribution is. Usually, we just need the ‘effect size’



Z-test power functions

- Get the power given *d*, *n*, and *alpha*. (2-tailed!)

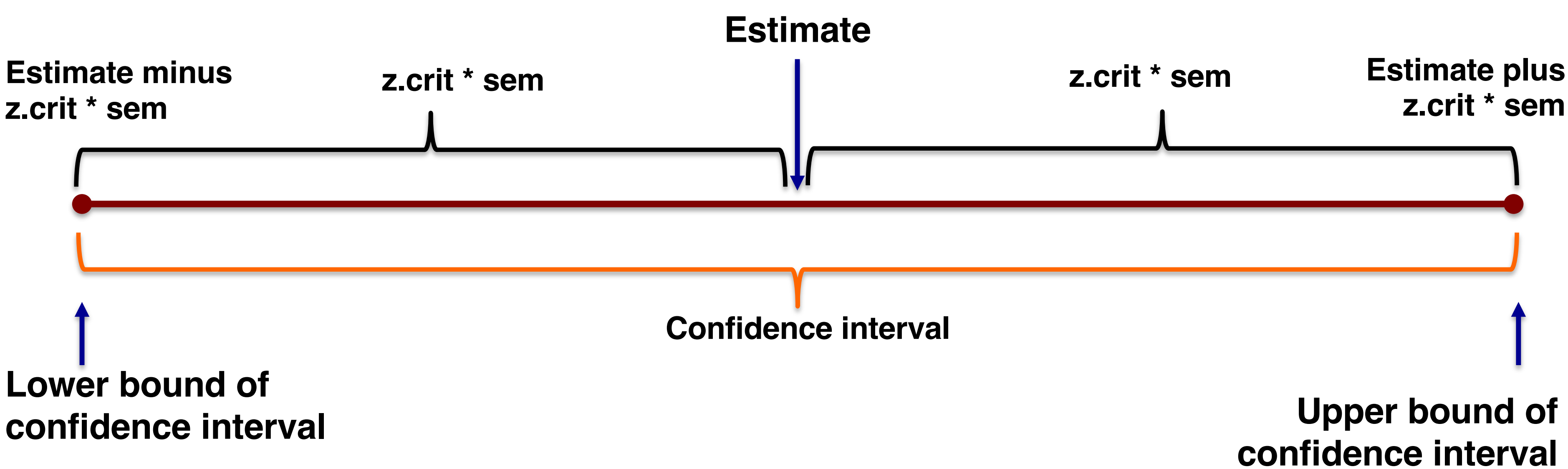
```
pwr::pwr.norm.test(d=d, n=n, sig.level=alpha)
```

- Get the necessary n to reach *power*, given *d*, and *alpha.*

```
pwr::pwr.norm.test(d=d, sig.level=alpha, power=power)
```

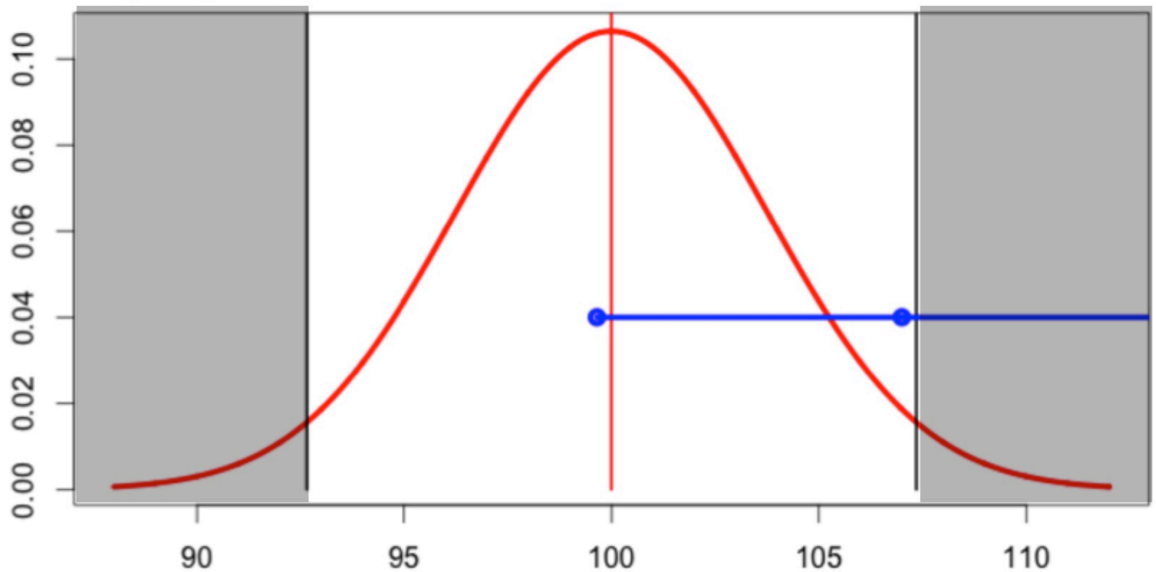
q=(1-α)% confidence interval on mean

$$\bar{x} \pm z_{\alpha/2}\sigma_0 / \sqrt{n}$$

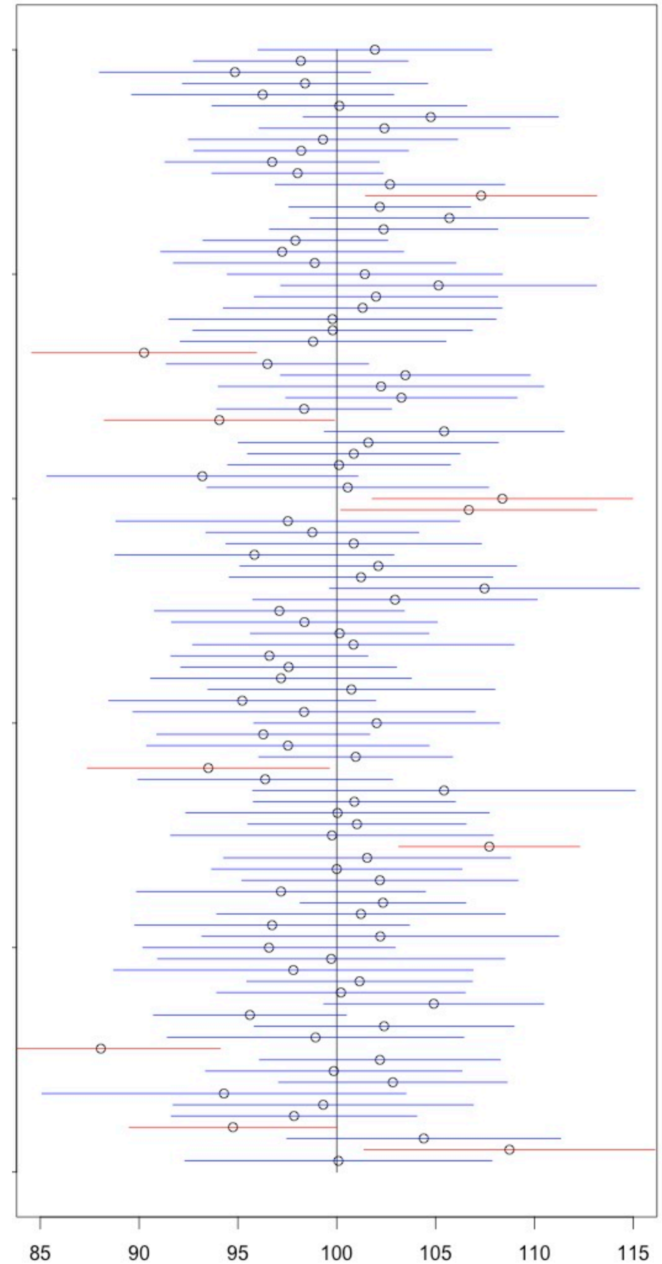


Confidence intervals

- If a 90% confidence interval on the mean excludes the null hypothesis mean, we can reject that null hypothesis with 2-tailed alpha = 0.1, and vice versa.



- We expected 90 out of 100 90% confidence intervals to include the true mean.
  "90%" refers to a long-run property of the procedure used to define the confidence interval, not to the specific confidence interval you have.



Probabilities in classical statistics refer to sampling frequencies under some statistical model.

- p-value: what proportion of hypothetical samples from the *null hypothesis model*, would have a statistic at least as extreme as ours?
- Alpha: probability of rejecting the null hypothesis for data sampled from the *null hypothesis model*.
- Power: probability of rejecting the null hypothesis for data sampled from some *alternative model*.
- Sampling distribution: the probability distribution of a statistic given that it is sampled from *some model*.
- Confidence interval probability: probability that a confidence interval computed in this manner using samples from *some model* will contain the model parameter value.

Probabilities in null hypothesis significance testing refer to peculiar conditional probabilities:

- p-value:
  P(X › x.sample | null is true)   P(X › x.sample | X~null)

- Alpha:
  P(significant | null is true)

- Power:
  P(significant | null is false)

- *Really important:*
  - *These do not give us the probability that the null is false: P(null is false | significant) !!*

Today

- T-tests: why, how, varieties.

- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

Normal variable stats.

- NHST: Z-test.
  - Get (2-tailed) p-value via

$$z_{\bar{x}} = \frac{\bar{x} - \mu_0}{\sigma_X}\sqrt{n}$$

```
2*pnorm(-abs(z),0,1)
```

- Confidence intervals on mean
  - Equivalence to NHSTs!

$$\bar{x} \pm z_{\alpha/2}\sigma_0 / \sqrt{n}$$

```
za = qnorm(a/2,0,1)
```

- Effect size
  - Scale and sample size neutral.

$$d = \left| \frac{\mu_T - \mu_0}{\sigma_X} \right|$$

- Alpha, Beta, Power.
  - Effect size and n matter.

```
pow    =   1 - pnorm(  abs(qnorm(a/2)) - d*sqrt(n)  )
n.needed   =  ( ( qnorm(a/2) - qnorm(pow) ) / d )^2
```

```
pwr::pwr.norm.test
```

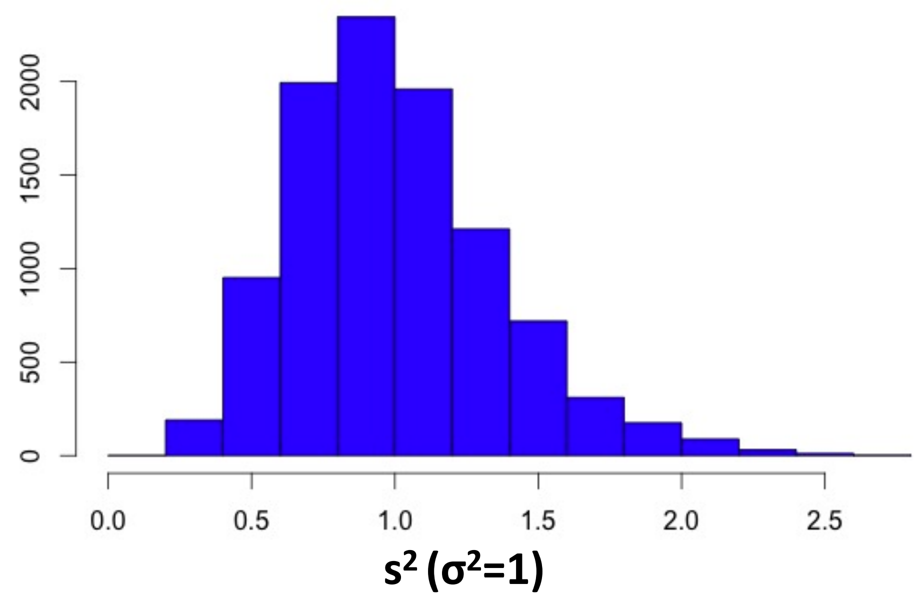
- Critically: Known standard deviation? What if our Ho just specifies the *mean*

Sample standard deviation varies

s^2 (sample variance) has sampling variation

`var(rnorm(16,0,1))`	[1] 0.748
`var(rnorm(16,0,1))`	[1] 0.966
`var(rnorm(16,0,1))`	[1] 0.830
`var(rnorm(16,0,1))`	[1] 1.292

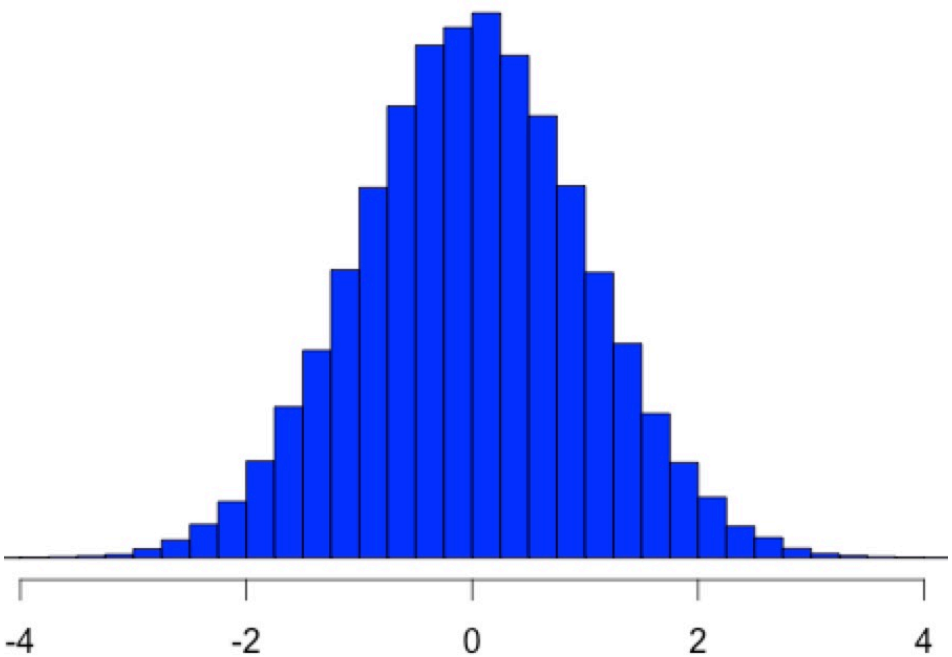
```
replicate(10000,var(rnorm(16,0,1)))
```

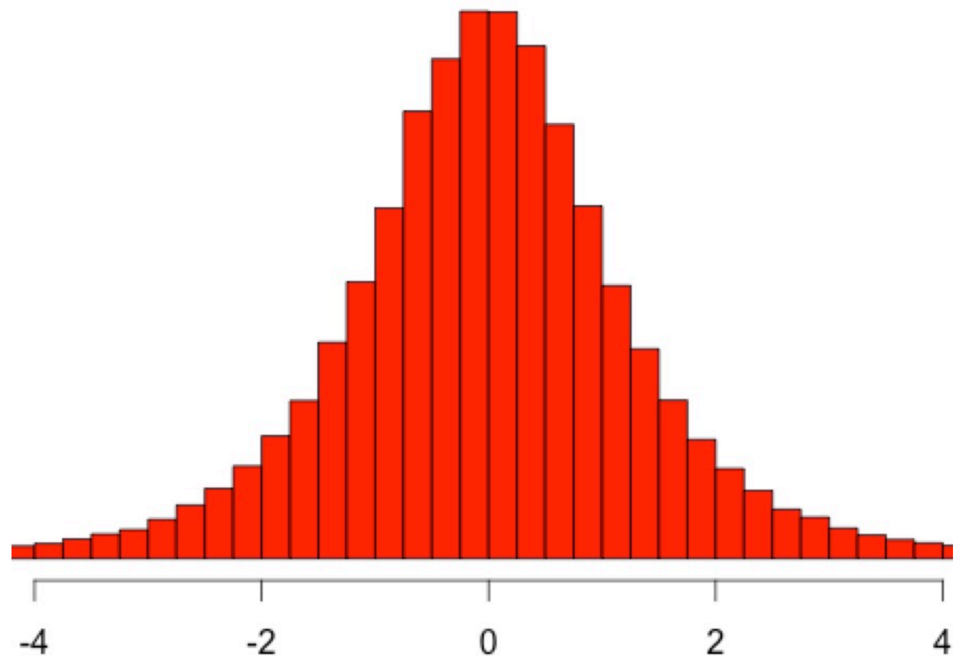


Null distribution of Z and T statistic

$$z_{\bar{x}} = \left(\frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \right)$$

$$t_{\bar{x}} = \left(\frac{\bar{x} - \mu_0}{s_{\bar{x}}} \right)$$





Heavier tails in t statistic due to sampling variability of sample s.d.
Consequently, we use t-distribution. (pt, qt, rt, dt)

Z vs T statistic (one sample mean)

$$\sigma_{\bar{x}} = \sigma_0 / \sqrt{n}$$

Use sample sd. instead of null population sd. to define standard error

$$s_{\bar{x}} = s_x / \sqrt{n}$$

$$z_{\bar{x}} = \left(\frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} \right)$$

$$t_{\bar{x}} = \left(\frac{\bar{x} - \mu_0}{s_{\bar{x}}} \right)$$

Normal dist. equal to t dist. with df=infinity

Degrees of freedom

$$df = n - 1$$

```
2*pnorm(-abs(a))
qnorm(alpha/2)
```

pt and qt instead of pnorm and qnorm

(1-alpha)% confidence interval

```
2*pt(-abs(t),df)
qt(alpha/2,df)
```

Varieties of t-tests

Testing / confidence intervals using sample std. devs.

- Is the mean math GRE score of psych students different from 700?

  "One-sample" t-test

- Is the avg. math GRE score for psych students different from cog sci students?

  "Two-sample" t-test (perhaps equal variance.)

- Is the avg. improvement in math GRE scores from taking a Kaplan course different from 0?

  "Paired sample" t-test (one-sample t-test on difference)

- Is the avg. improvement from taking a Kaplan course different from the avg. improvement from just taking a bunch of practice GREs?

  "Two-sample" t-test (after calc. deltas, perhaps unequal variance?)

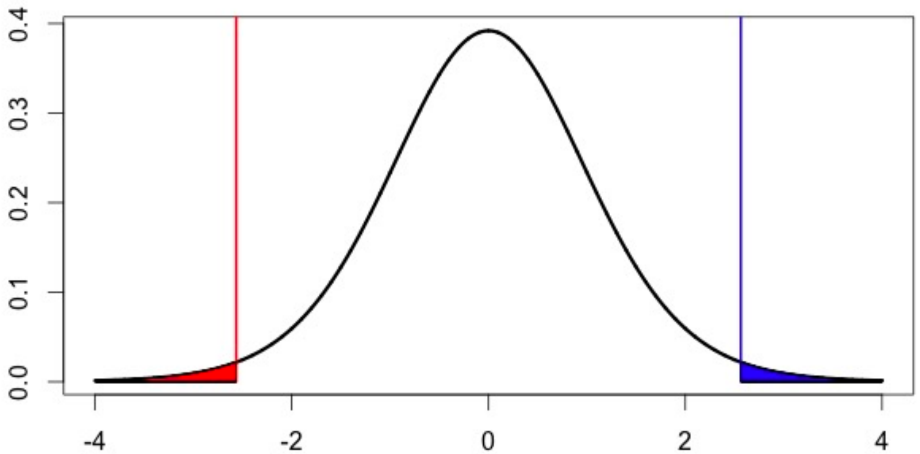
One sample t-test

We have a sample from population with unknown variance, and we want to know if the mean of that population is different from some H0 mean.

Is the mean math GRE score of psych students different from 700?

```
x = c(618,606,735,627,679,622,712,772,728,550,594,681,578,689,672)
```

```
t.test(x, mu=700)

     One Sample t-test

data:  x
t = -2.5645, df = 14, p-value = 0.02248
alternative hypothesis: true mean is not equal
to 700
95 percent confidence interval:
 622.0167 693.0500
sample estimates:
mean of x
 657.5333
```



Lower tail p-val (0.0112: 1-tail p-val)

The other tail (0.0112) for 2-tail test.

Two sample t-test (assumed equal variance)

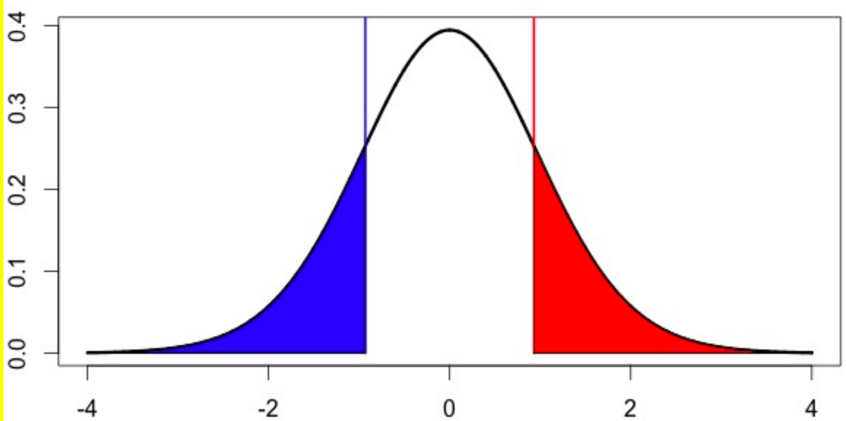
We have samples from two population with unknown variance (but equal variance), and we want to know if their population means are different from each other.

Is the avg. math GRE score for psych students different from cog sci students?

```
x1 = c(618,606,735,627,679,622,712,772,728,550,594,681,578,689,672)
x2 = c(571,569,613,693,714,521,530,736,677,626,722)
```

```
t.test(x1,x2,var.equal=TRUE)
```

```
     Two Sample t-test

data:  x1 and x2
t = 0.8458, df = 24, p-value = 0.406
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -34.15577  81.58608
sample estimates:
mean of x mean of y
 657.5333  633.8182
```



Paired sample t-test (one-sample on differences)

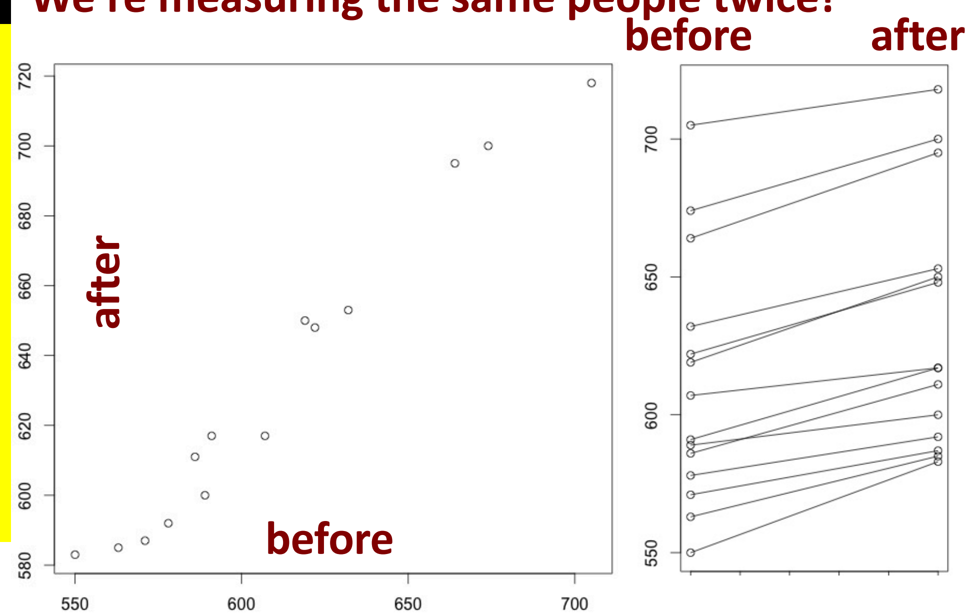
Is the avg. improvement in math GRE scores from taking a Kaplan course different from 0?

Before:
```
xb = c(586,589,571,705,550,632,674,664,578,563,619,607,591,622)
```
After:
```
xa = c(611,600,587,718,583,653,700,695,592,585,650,617,617,648)
```

```
t.test(xb, xa, var.equal=TRUE)
```

```
     Two Sample t-test

data:  xb and xa
t = -1.2691, df = 26, p-value = 0.2157
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -57.07210  13.50068
sample estimates:
mean of x mean of y
 610.7857  632.5714
```

We're measuring the same people twice!



And individuals seem to be improving…

Paired sample t-test (one-sample on differences)

We have two measurements of the same ‘subjects’ from the population, and we want to know if there was a change.

Is the avg. improvement in math GRE scores from taking a Kaplan course different from 0?

Before:
```
xb = c(586,589,571,705,550,632,674,664,578,563,619,607,591,622)
```
After:
```
xa = c(611,600,587,718,583,653,700,695,592,585,650,617,617,648)
```

Strategy: factor out the across-person variation by looking at the *change* within person.

```
D = xa-xb
```
```
[1] 25 11 16 13 33 21 26 31 14 22 31 10 26 26
```
changes

```
t.test(D)
```
```
     One Sample t-test

data:  D
t = 10.4809, df = 13, p-value = 1.041e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 17.29514 26.27629
sample estimates:
mean of x
 21.78571
```

Paired sample t-test is just a one-sample t-test with a sample of the *differences*!

This allows us to factor our across-person variation, which makes such *repeated measures* designs/tests very powerful!

Two sample t-test (unequal variance)

We have samples from two population with unknown (but potentially unequal) variance, and we want to know if their population means are different from each other.

Is the avg. improvement from taking a Kaplan course different from the avg. improvement from just taking a bunch of practice GREs?

```
xD = c(25,11,16,13,33,21,26,31,14,22,31,10,26,26)
yD = c(-9,-19,16,18,46,8,30,45,25,33,11,5,23,22,38,32,-2)
```

Kaplan improvement

Regular improvement

```
t.test(xD, yD)
```

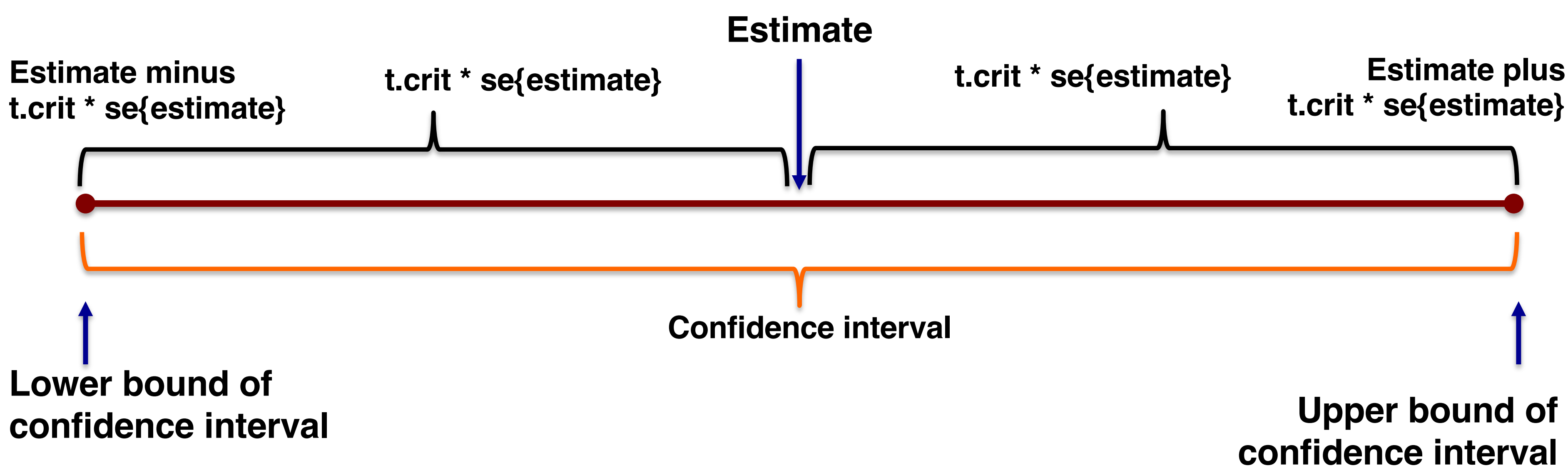
```
     Welch Two Sample t-test

data:  xD and yD
t = 0.5797, df = 22.443, p-value = 0.5679
alternative hypothesis: true difference in means is
not equal to 0
95 percent confidence interval:
 -7.319357 13.008433
sample estimates:
mean of x mean of y
 21.78571  18.94118
```

Confidence intervals.

Our confidence intervals are of this form:

Estimate — Critical standard score — Standard error of the estimate

$$\hat{\theta} \pm t_{\alpha/2} \cdot s\{\hat{\theta}\}$$



Confidence intervals.

Estimate:

sample mean (1-sample test)
difference between sample means (2-sample test)
sample mean of differences (paired test)

Critical score:

q = percent of interval (e.g., 0.9); alpha = 1-q

```
t.crit = abs( qt( (1-q)/2 , df ) )
```

Standard error of estimate:

matched to estimate, derived from expectation...

```
estimate + c(-1,1) * t.crit * se.estimate
```

	Z-test	One-sample t-test	Paired t-test	2-sample eq. var. t-test	2-sample uneq var t-test
	We know pop. var. Want to test if mean differs from H0 mean.	We **do not know** pop. var. Want to test if mean differs from H0 mean.	We have 2 measures of the same thing, do they differ in means?	We want to know if two samples (assumed to have equal var) have different means	We want to know if two samples (not assumed to have equal var) have different means
Statistic	$z_{\bar{x}} = \left(\frac{\bar{x}-\mu_0}{\sigma_X}\right)\sqrt{n}$	$t_{\bar{x}} = \left(\frac{\bar{x}-\mu_0}{s_X}\right)\sqrt{n}$	$t_{\bar{D}} = \left(\frac{\bar{D}}{s_D}\right)\sqrt{n}$	$t_{\bar{x}-\bar{y}} = \frac{\bar{x}-\bar{y}}{s_P\sqrt{\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$	$t_{\bar{x}-\bar{y}} = \frac{\left(\bar{x}-\bar{y}\right)}{\sqrt{\left(\frac{s_X^2}{n_x}+\frac{s_Y^2}{n_y}\right)}}$
Effect size	$\hat{d} = \left(\frac{\bar{x}-\mu_0}{\sigma_X}\right)$	$\hat{d} = \left(\frac{\bar{x}-\mu_0}{s_X}\right)$	$\hat{d} = \left(\frac{\bar{D}}{s_D}\right)$	$\hat{d} = \left(\frac{(\bar{x}-\bar{y})-\mu_0}{s_P}\right)$	Effect size here breaks the mold because of the diff. variances.
d.f.		$df = n-1$	$df = n-1$	$df = n_1+n_2-2$	$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$
P-value 2-tailed	`2*pnorm(-abs(z))`	`2*pt(-abs(t),df)`			
1-α% C.I.	$\bar{x} \pm z_{\alpha/2}\boxed{\sigma_X/\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2}\boxed{s_X/\sqrt{n}}$	$\bar{D} \pm t_{\alpha/2}\boxed{s_D/\sqrt{n}}$	$(\bar{x}-\bar{y}) \pm t_{\alpha/2}\boxed{s_P * \sqrt{1/n_1+1/n_2}}$	$(\bar{x}-\bar{y}) \pm t_{\alpha/2}\boxed{\sqrt{\frac{s_X^2}{n_x}+\frac{s_Y^2}{n_y}}}$
	`z* = qnorm(a/2)`	`t* = qt(a/2,df)`			

Standard errors of the mean / difference

T-test power

```
library(pwr)

pwr.t.test(n = 30, d = 0.5, sig.level=0.05, type="two.sample")

pwr.t.test(d = 0.5, sig.level=0.05, power=0.8, type="two.sample”)
```

Today

- T-tests: why, how, varieties.

- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

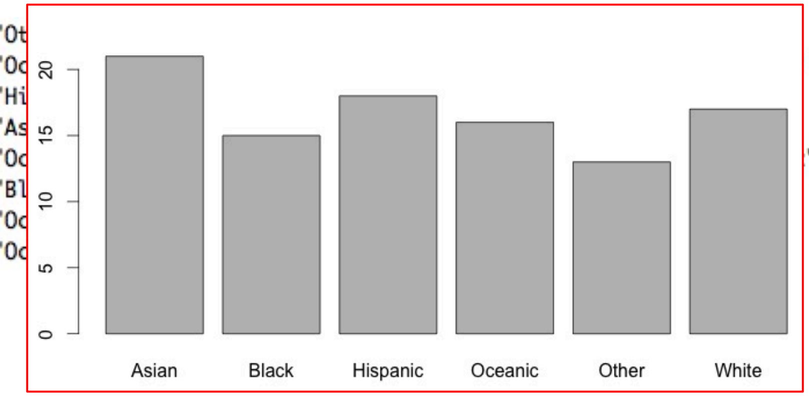
Counts vs Categories

- ‘Count data’:
  You count the number of occurrences, and you do not keep track of the non-occurrences.
  - Number of cars in parking lot.
  - Number of people entering a store
  - Number of spikes in a post-stimulus interval
  - Number of babies in an expt
  - Number of voters at a given polling place

- ‘Categorical data’:
  Each observation is categorized into one of several mutually exclusive labels. Every observation goes into exactly one bin.
  - Makes of cars in parking lot.
  - Race of people entering a store.
  - Types of cells the spikes came from.
  - ASD categorization of participating babies
  - Who those people voted for

Describing categorical data...

```
 [1] "White"    "Other"    "Asian"    "Oceanic"  "Other"    "Hispanic" "Hispanic" "Ot
[14] "Other"    "Oceanic"  "Asian"    "White"    "Asian"    "Hispanic" "Hispanic" "Oc
[27] "Hispanic" "Asian"    "Oceanic"  "Black"    "Asian"    "Oceanic"  "Asian"    "Hi
[40] "Asian"    "Black"    "Oceanic"  "Oceanic"  "Asian"    "Hispanic" "Other"    "As
[53] "White"    "Hispanic" "Black"    "White"    "Other"    "Hispanic" "Oceanic"  "Oc
[66] "Hispanic" "Hispanic" "Black"    "Asian"    "Black"    "Black"    "Oceanic"  "Bl
[79] "Hispanic" "White"    "Hispanic" "Hispanic" "Hispanic" "Other"    "Other"    "Oc
[92] "Hispanic" "White"    "Oceanic"  "Black"    "Other"    "White"    "Asian"    "Oc
```



- Categorize and count

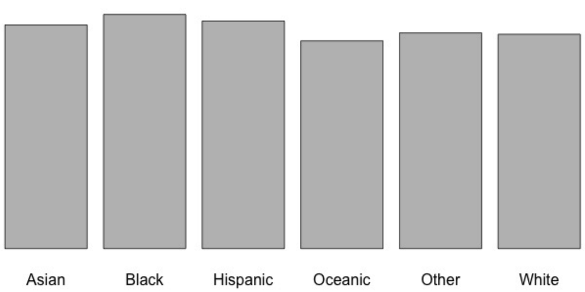
```
table(x)
```

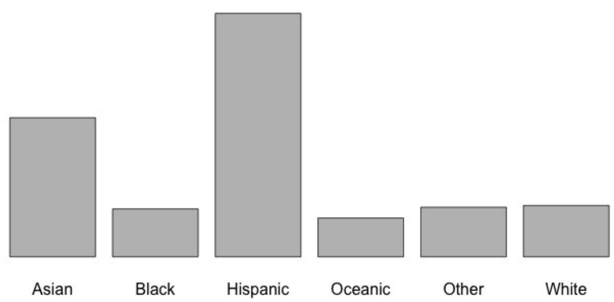
```
   Asian    Black Hispanic  Oceanic    Other    White
      21       15       18       16       13       17
```

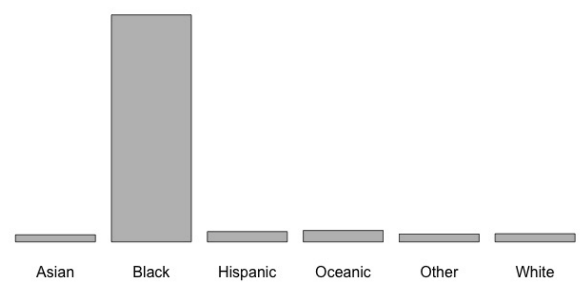
- Estimate mode: ‘Asian’

```
z = table(x)
names(z)[which.max(z)]
```

- Dispersion as... not-peakiness (entropy)?
  (not a standard measure, but may be useful)







Today

- T-tests: why, how, varieties.

- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

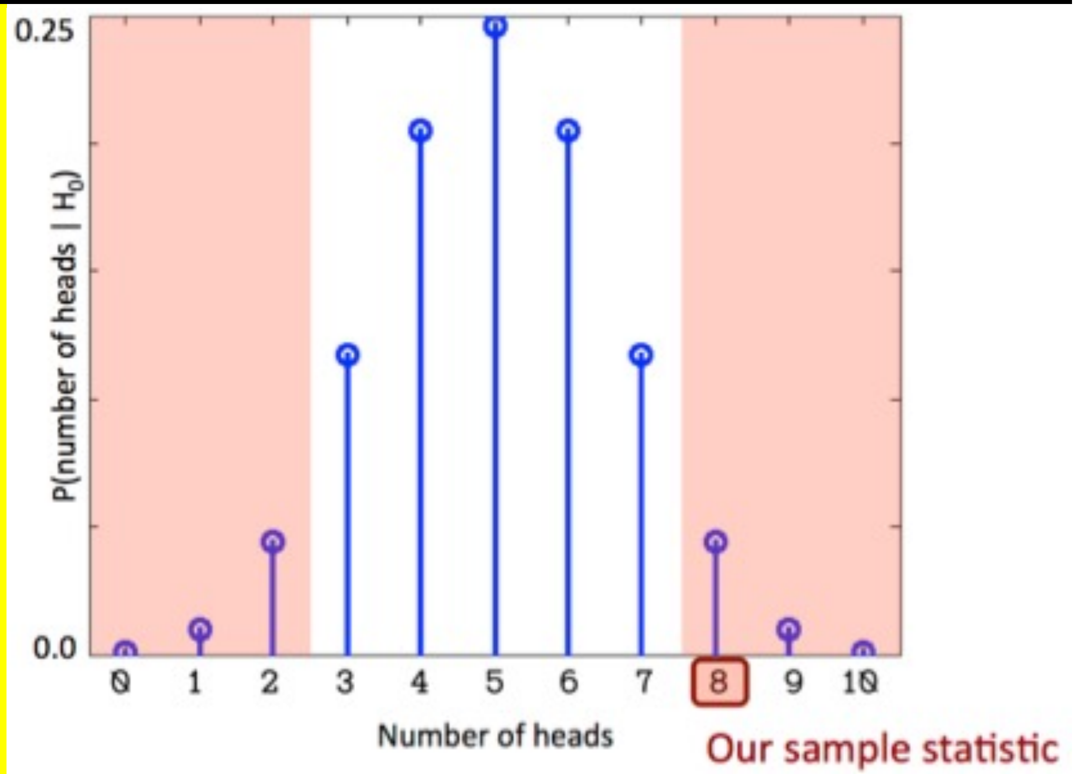
- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

Binomial test

- Having seen *k* "successes" out of *n* attempts, can we reject the null of binomial draws with probability *p*.
  - Boys/total born in hospital, correct/total problems, etc.
- Binomial test: compare k to binomial with n, p

```
binom.test(x=8, n=10, p=0.5)
```

```
     Exact binomial test

data:  8 and 10

number of successes = 8,
number of trials = 10,
p-value = 0.1094

alternative hypothesis:
     true probability of success
     is not equal to 0.5
95 percent confidence interval:
 0.4439045 0.9747893
sample estimates:

probability of success
                   0.8
```



Sign test (Binomial test for quantile)

- We have numerical data and want to test a null median.
  - Why? Because data have weird distribution (skew, kurtosis, outliers), so a t-test for mean is weak given large s.d.
  - Most often: difference before-after scores to test for median difference of zero.
  - Logic: if median is M, then $P(x>M)=0.5$.

```
x    1.04  0.82  0.79  1.08  0.71 -1.39 -2.61  1.24 10.84 -2.94  0.87  0.80  0.48  2.82  1.75
t.test(x,mu=0)$p.value                                                                  0.197

binom.test(x = sum(x>0), n=sum(x>0 | x<0), p=0.5)
data:  sum(x > 0) and sum(x > 0 | x < 0)

number of successes = 12,
number of trials = 15,
p-value = 0.03516
```

Note:
we do not count x values *exactly* equal to the median!

- Same logic applies to other quantiles, not often used.

Normal approximation of Binomial p.

- p.hat = k/n
  (proportion of successes out of number of attempts)
  if n is big enough sampling distribution of p.hat will…
  …be normally distributed (Central limit theorem)
  …have a mean of p (an *unbiased* estimate)
  …have s.d.: **sqrt(p*(1-p) / n)** <- "standard error!"
- Consequently, p.hat ~ Normal(p, sqrt(p*(1-p)/N))
  and we can use the logic of Z confidence intervals.
- Confidence interval on Binomial p:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

```
p.hat = k/n
se.p.hat = sqrt(p.hat*(1-p.hat)/n)
p.hat + c(1,-1)*qnorm((1-q)/2)*se.p.hat
```

Binomial tests, proportions

- Hospital gets 30 boys out of 50 births.
  - Can we reject null hypothesis of 50% girls?
  - What is the 95% confidence interval on the proportion of boys born in that hospital?

Today

- T-tests: why, how, varieties.

- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

Chi-squared goodness of fit.

- Categorical data in $c > 2$ categories.
- Distilled into counts $k_1, k_2, \ldots, k_c$.
- Test null category probabilities: $p_1, p_2, \ldots p_c$

Consider SPSP membership ethnicities:

```
spsp = read.csv(
            url('http://vulstats.ucsd.edu/data/spsp.demographics.tsv'),
            sep='\t')
spsp$ethnicity = as.character(spsp$ethnicity)
spsp$ethnicity[spsp$ethnicity==""] = "No Report"
str(spsp)
```

```
'data.frame':    5694 obs. of  3 variables:
 $ stage    : Factor w/ 5 levels "Early Career",..: 5 5 5 5 5 5 ...
 $ gender   : Factor w/ 4 levels "","Female","Male",..: 2 2 2 2 ...
 $ ethnicity: chr  "Black" "Black" "Black" "Black" ...
```

Chi-squared goodness of fit.

Categorical data in $c > 2$ categories.

Distilled into counts $k_1, k_2, \ldots, k_c$.

```
unique(spsp$ethnicity)
"Black"
"Native American"
"Asian"
"White"
"Latino"
"Arab"
"Other"
"No Report"
```

```
table(spsp$ethnicity)
Arab               23
Asian             657
Black             156
Latino            151
Native American    36
No Report         146
Other            1308
White            3217
```

Test null category probabilities: $p_1, p_2, \ldots p_c$

Wait… what is the null hypothesis? (uh.. let's say US dist.)

```
null.p = c("Latino"=0.164,
           "White"=0.637,
           "Asian"=0.047,
           "Black"=0.122,
           "Native American"=0.007,
           "Other"=0.019+0.002+0.002)
```

Complication:
We need to make data categories match these categories…

Cleaning up data for chi.squared fx.

Categorical data in $c > 2$ categories.

```
unique(spsp$ethnicity)
```

Distilled into counts $k_1, k_2, \ldots, k_c$.

```
K=table(spsp$ethnicity)
```

Test null category probabilities: $p_1, p_2, \ldots p_c$

```
null.p
```

Make categories in data match categories in null.

```
K.t = c("Latino"=K[['Latino']],
         "White"=K[['White']],
         "Asian"=K[['Asian']],
         "Black"=K[['Black']],
         "Native American"=K[['Native American']],
         "Other"=K[['Arab']]+K[['Other']])
```

Make sure order of null.p and K.t matches.

```
c.order = sort(names(null.p))
null.p = null.p[c.order]
K.t = K.t[c.order]
```

Chi-squared goodness of fit.

Categorical data in $c > 2$ categories.

Distilled into counts $k_1, k_2, \ldots, k_c$.

`K.t`

Test null category probabilities: $p_1, p_2, \ldots p_c$

`null.p`

Test for significant deviation of counts from null probs.

```
chisq.test(x=K.t, p=null.p)
```

```
        Chi-squared test for given probabilities

data:  K.t
X-squared = 13013, df = 5, p-value < 2.2e-16
```

What is a chi-squared statistic?

sum((observed – expected)²/expected)

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

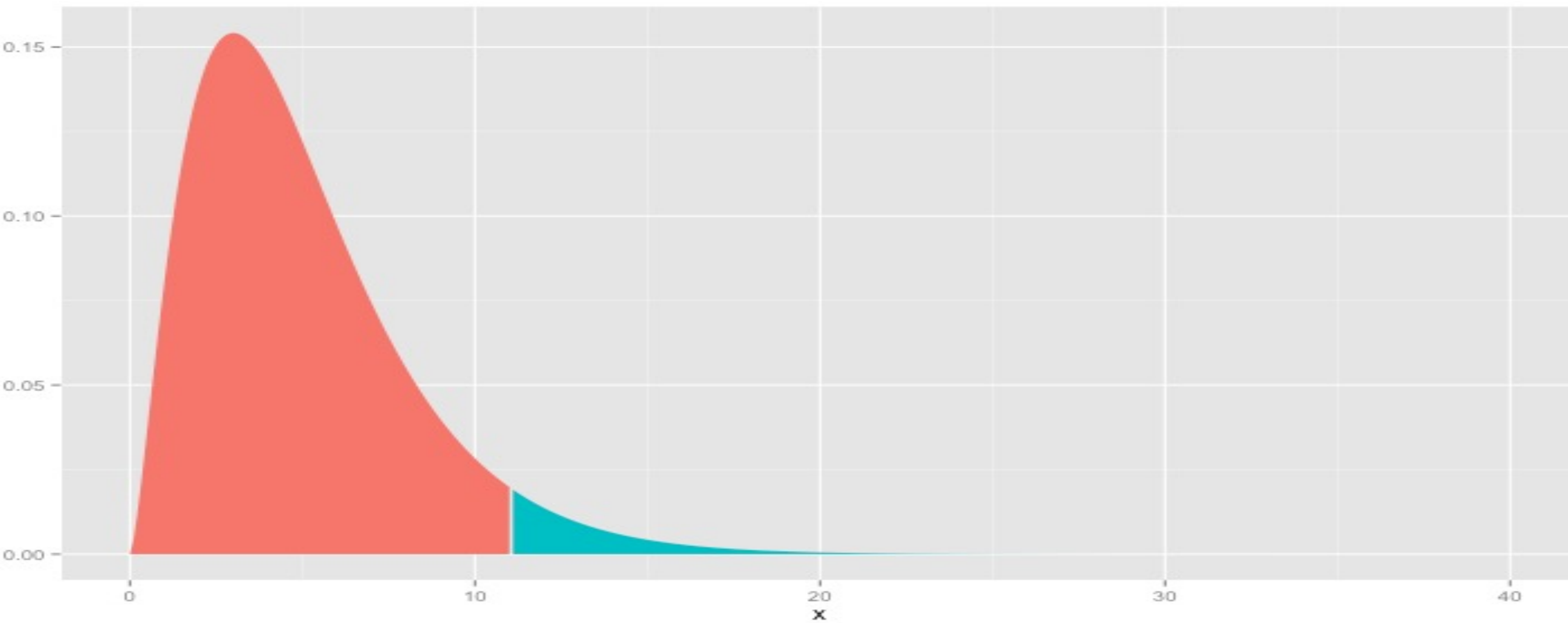
	Asian	Black	Latino	Nat. Am	Other	White	TOTAL
Observed	657	156	151	36	1331	3217	5548
Null.p	0.047	0.122	0.164	0.007	0.023	0.637	
Expected	260.8	676.9	909.9	38.8	127.6	3534.1	
obs-exp	396.2	-520.9	-758.9	-2.8	1203.4	-317.1	
(o-e)^2	157009.3	271291.0	575886.7	8.0	1448161.9	100537.2	
(o-e)^2/e	602.1	400.8	632.9	0.2	11348.9	28.4	**13013.4**

df = c-1

(number of categories – 1. relationship to degrees of freedom in t dist.)

```
p.value = 1-pchisq(13013,df)        0
```

Chi-squared distribution

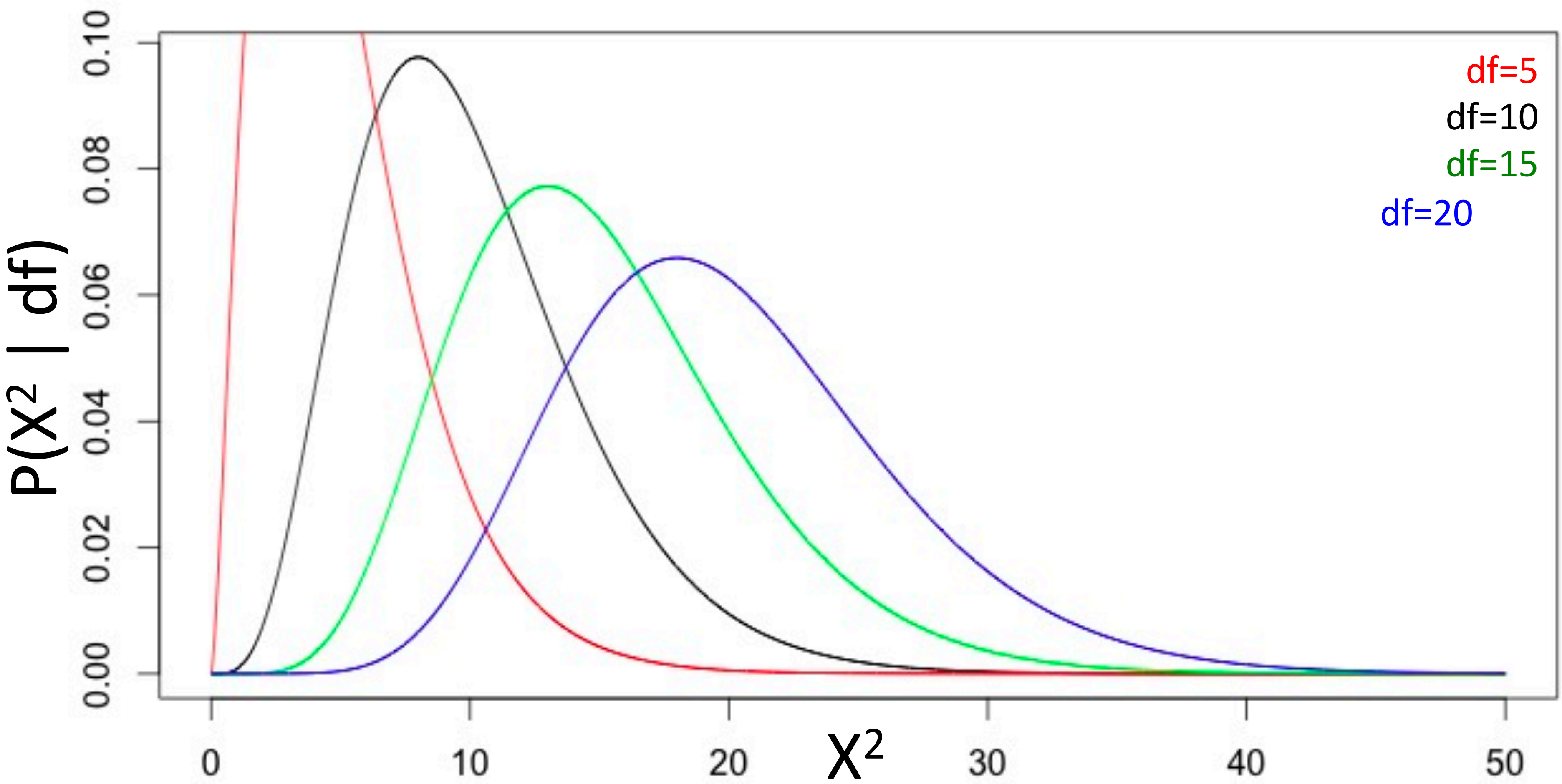


To reject the null, we do a one-tailed test: since only positive values constitute a large deviation from the null.

```
p.value = 1-pchisq(X2,df)        0
```

Small deviations indicate data *too* consistent with the null

Chi-squared distribution



Chi-squared test for null category ps

Categorical data in $c > 2$ categories.

Distilled into counts $k_1, k_2, \ldots, k_c$.

`K.t`

Test null category probabilities: $p_1, p_2, \ldots p_c$

`null.p`

Test for significant deviation of counts from null probs.

```
chisq.test(x=K.t, p=null.p)
```

```
	Chi-squared test for given probabilities

data:  K.t
X-squared = 13013, df = 5, p-value < 2.2e-16
```

Today

- T-tests: why, how, varieties.

- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

Contingency table independence tests

Categorize data into two orthogonal categorical variables

Result: a C x R “contingency” table of counts.

```
observed = table(spsp$stage, spsp$ethnicity)
```

	Arab	Asian	Black	Latino	Native American	No Report	Other	White
Early Career	4	57	14	7	4	20	108	273
Grad	13	328	68	63	20	58	515	1098
Regular Member	4	227	54	55	9	55	549	1557
Retired	0	1	1	0	1	7	28	86
Undergrad	2	44	19	26	2	6	108	203

Null: category variables are independent, meaning

$$P(col,row)=P(col)*P(row)$$

```
chisq.test(observed)
```

```
                Pearson's Chi-squared test
data:  observed
X-squared = 156.86,
df = 28,
p-value < 2.2e-16

Warning message:
In chisq.test(observed) : Chi-squared approximation may be incorrect
```

What’s this warning?
More on this later

Chi-squared independence calculation

```
observed = table(spsp$stage, spsp$ethnicity)
```

```
n = sum(observed)
```

5694

```
p.row = rowSums(observed)/n
```

```
Early Career     Grad Regular Member   Retired   Undergrad
       0.086    0.380          0.441     0.022       0.072
```

```
p.col = colSums(observed)/n
```

```
 Arab Asian Black Latino Native American  No Report  Other  White
0.004 0.115 0.027  0.027           0.006      0.026  0.230  0.565
```

```
p.indep = outer(p.row,p.col,function(pc,pr)(pr*pc))
```

```
                Arab Asian Black Latino Native American No Report Other White
Early Career   0.000 0.010 0.002  0.002           0.001     0.002 0.020 0.048
Grad           0.002 0.044 0.010  0.010           0.002     0.010 0.087 0.215
Regular Member 0.002 0.051 0.012  0.012           0.003     0.011 0.101 0.249
Retired        0.000 0.003 0.001  0.001           0.000     0.001 0.005 0.012
Undergrad      0.000 0.008 0.002  0.002           0.000     0.002 0.017 0.041
```

Chi-squared independence calculation

```
observed = table(spsp$stage, spsp$ethnicity)
```

```
n = sum(observed)
```

5694

```
expected = p.indep*n
```

	Arab	Asian	Black	Latino	Native American	No Report	Other	White
Early Career	1.97	56.19	13.34	12.91	3.08	12.49	111.87	275.15
Grad	8.74	249.58	59.26	57.36	13.68	55.46	496.87	1222.05
Regular Member	10.14	289.62	68.77	66.56	15.87	64.36	576.59	1418.10
Retired	0.50	14.31	3.40	3.29	0.78	3.18	28.48	70.06
Undergrad	1.66	47.31	11.23	10.87	2.59	10.51	94.18	231.64

```
(observed-expected)^2/expected
```

	Arab	Asian	Black	Latino	Native American	No Report	Other	White
Early Career	2.10	0.01	0.03	2.71	0.28	4.52	0.13	0.02
Grad	2.08	24.64	1.29	0.55	2.92	0.12	0.66	12.59
Regular Member	3.72	13.54	3.17	2.01	2.97	1.36	1.32	13.60
Retired	0.50	12.38	1.69	3.29	0.06	4.59	0.01	3.63
Undergrad	0.07	0.23	5.37	21.05	0.14	1.94	2.03	3.54

```
X2 = sum((observed-expected)^2/expected )
```

156.86

```
df = (nrow(observed)-1)*(ncol(observed)-1)
```

28

```
p.value = 1-pchisq(X2,df)
```

0

Degrees of freedom?

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

- Number of *unconstrained* elements that went into sum
- Number of elements minus the number of parameters.
  - n.row*ncol – (n.row-1) – (n.col-1) – 1
    # cells p.row p.col n
    5*8 – 4 – 7 – 1 = 28
  - Shortcut:
    (n.row-1)*(n.col-1) = 28

Contingency table independence tests

Categorize data into two orthogonal categorical variables

Result: a C x R “contingency” table of counts.

```
observed = table(spsp$stage, spsp$ethnicity)
```

```
                 Arab Asian Black Latino Native American No Report Other White
Early Career        4    57    14      7               4        20   108   273
Grad               13   328    68     63              20        58   515  1098
Regular Member      4   227    54     55               9        55   549  1557
Retired             0     1     1      0               1         7    28    86
Undergrad           2    44    19     26               2         6   108   203
```

Null: category variables are independent, meaning

$$P(col,row)=P(col)*P(row)$$

```
chisq.test(observed)
```

```
                 Pearson's Chi-squared test
data:  observed
X-squared = 156.86,
df = 28,
p-value < 2.2e-16

Warning message:
In chisq.test(observed) : Chi-squared approximation may be incorrect
```

What’s this warning?
More on this later

Theoretical vs. practical sampling dist.

- The X^2 statistic may not follow the X^2 distribution!
  - Only does so when the cell counts are sufficiently large for the ***Normal approximation to the binomial*** that underlies the statistic

```
chisq.test(x = c(10, 5, 3, 1), p = c(0.5, 0.25, 0.1, 0.15),
              simulate.p.value = F)
```

```
    Chi-squared test for given probabilities

data:  c(10, 5, 3, 1)
X-squared = 1.8772, df = 3, p-value = 0.5983

Warning message:
In chisq.test(c(10, 5, 3, 1), p = c(0.5, 0.25, 0.1, 0.15), simulate.p.value = F)
:
  Chi-squared approximation may be incorrect
```

```
chisq.test(x = c(10, 5, 3, 1), p = c(0.5, 0.25, 0.1, 0.15),
              simulate.p.value = T)
```

```
Chi-squared test for given probabilities with simulated p-value (based on 2000
replicates)

data:  c(10, 5, 3, 1)
X-squared = 1.8772, df = NA, p-value = 0.5892
```

Use the SPSP data…

```
spsp = read.csv(
url('http://vulstats.ucsd.edu/data/spsp.demographics.cleaned.csv’))
```

- Test for 50/50 male/female distribution among Grads, among Regular Members.
- Test for independence between male/female and Grad, Regular, and Undergrad.
- What is the 90% confidence interval on the proportion of White folks in the data set?
- Plot, somehow (in ggplot), the distribution of ethnicities as a function of “stage”

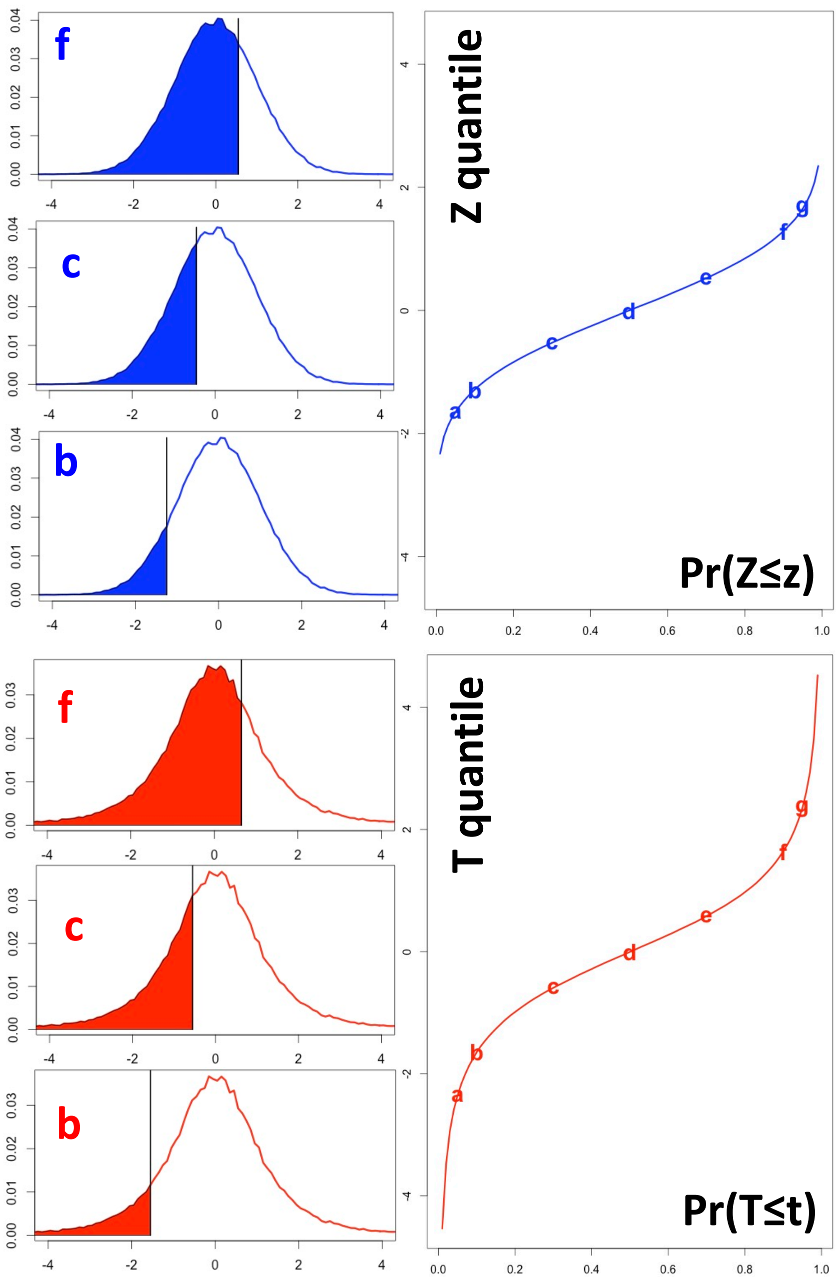
Today

- T-tests: why, how, varieties.

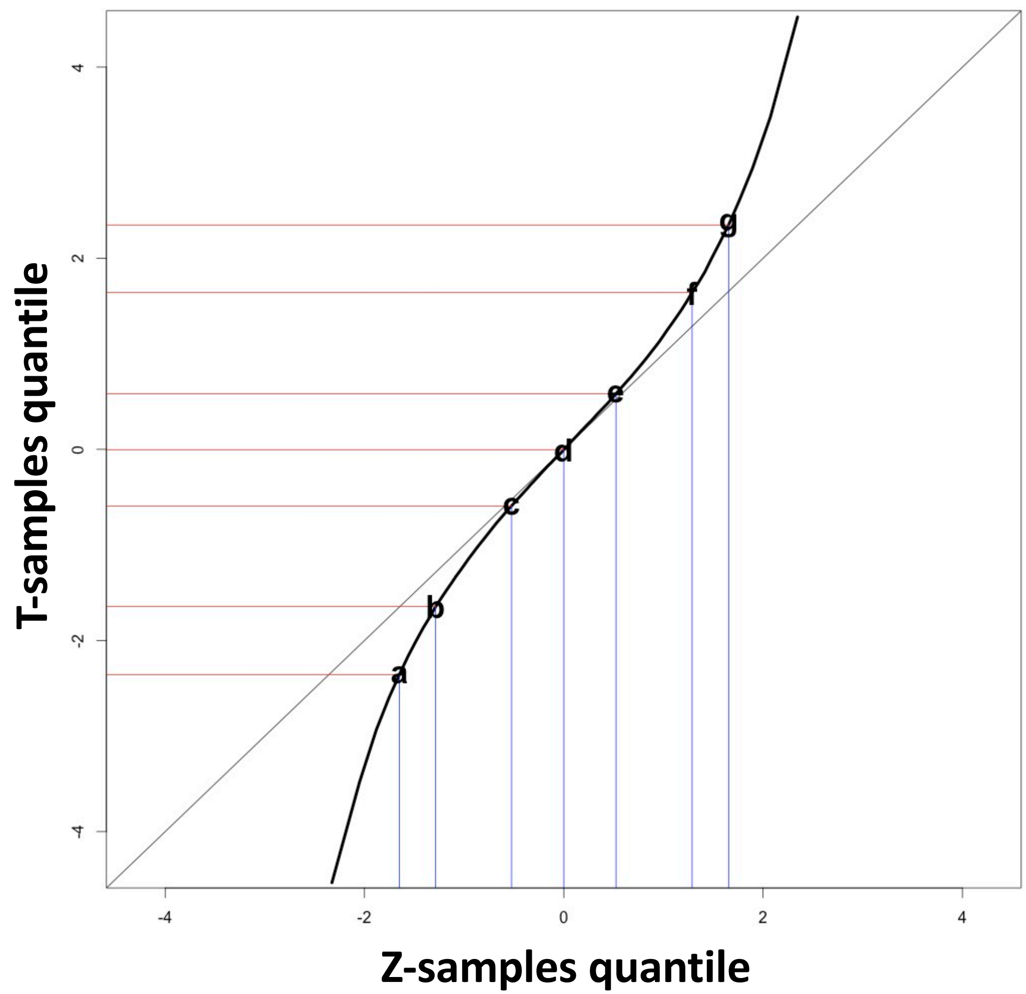
- Categorical data
  - Binomial proportions
  - Chi^2 goodness of fit
  - Chi^2 independence (for contingency tables)

- Optional (may not get to/cover)
  - QQ plots.
  - T-test formulas: working from summary statistics.
  - Standard errors: deriving.
  - What's up with df for unequal variance test?

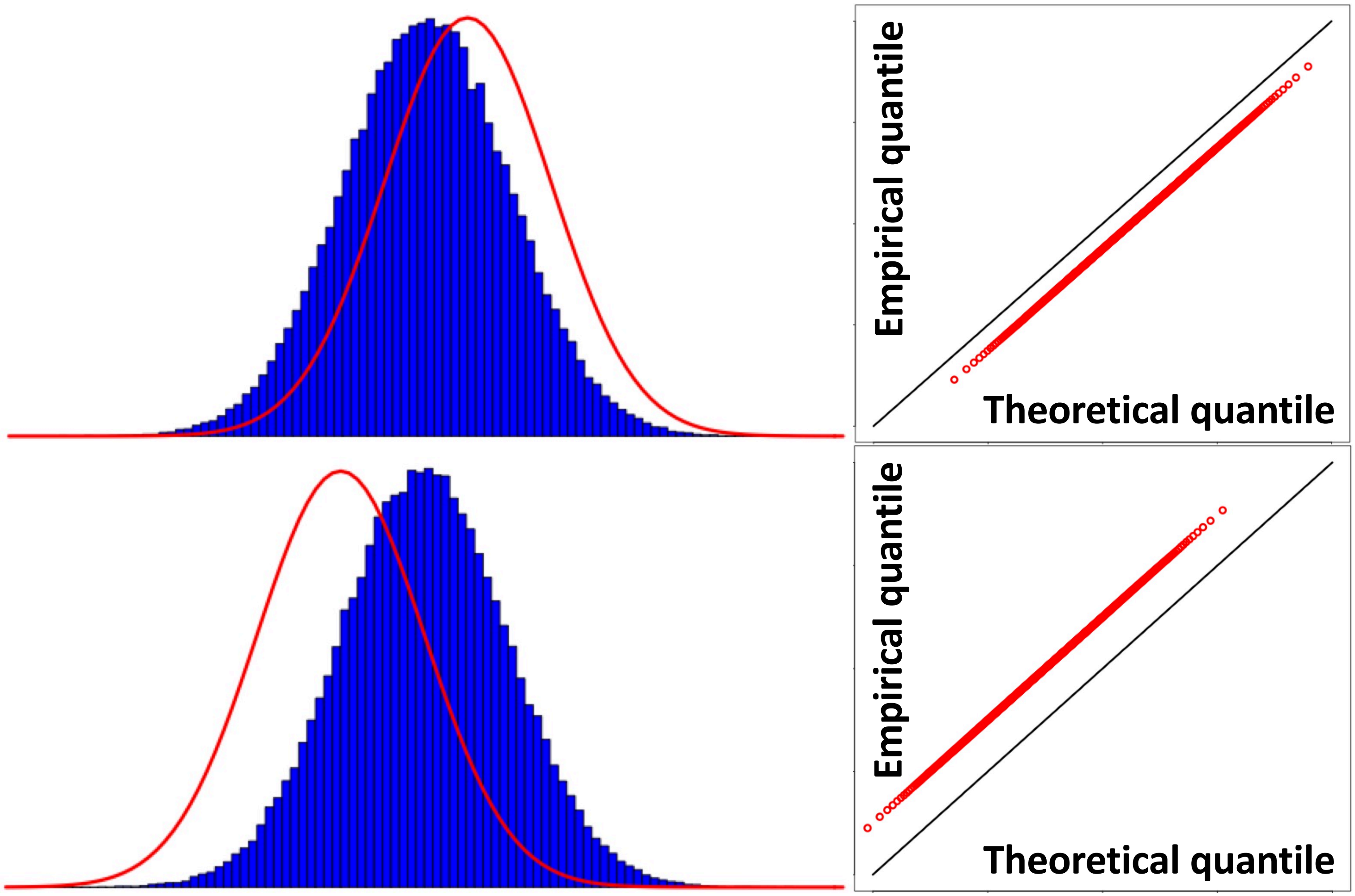
Q(uantile)-Q(uantile) plots



a: 0.05^{th} quantile
b: 0.10^{th} quantile
c: 0.30^{th} quantile
d: 0.50^{th} quantile
e: 0.70^{th} quantile
f: 0.90^{th} quantile
g: 0.95^{th} quantile

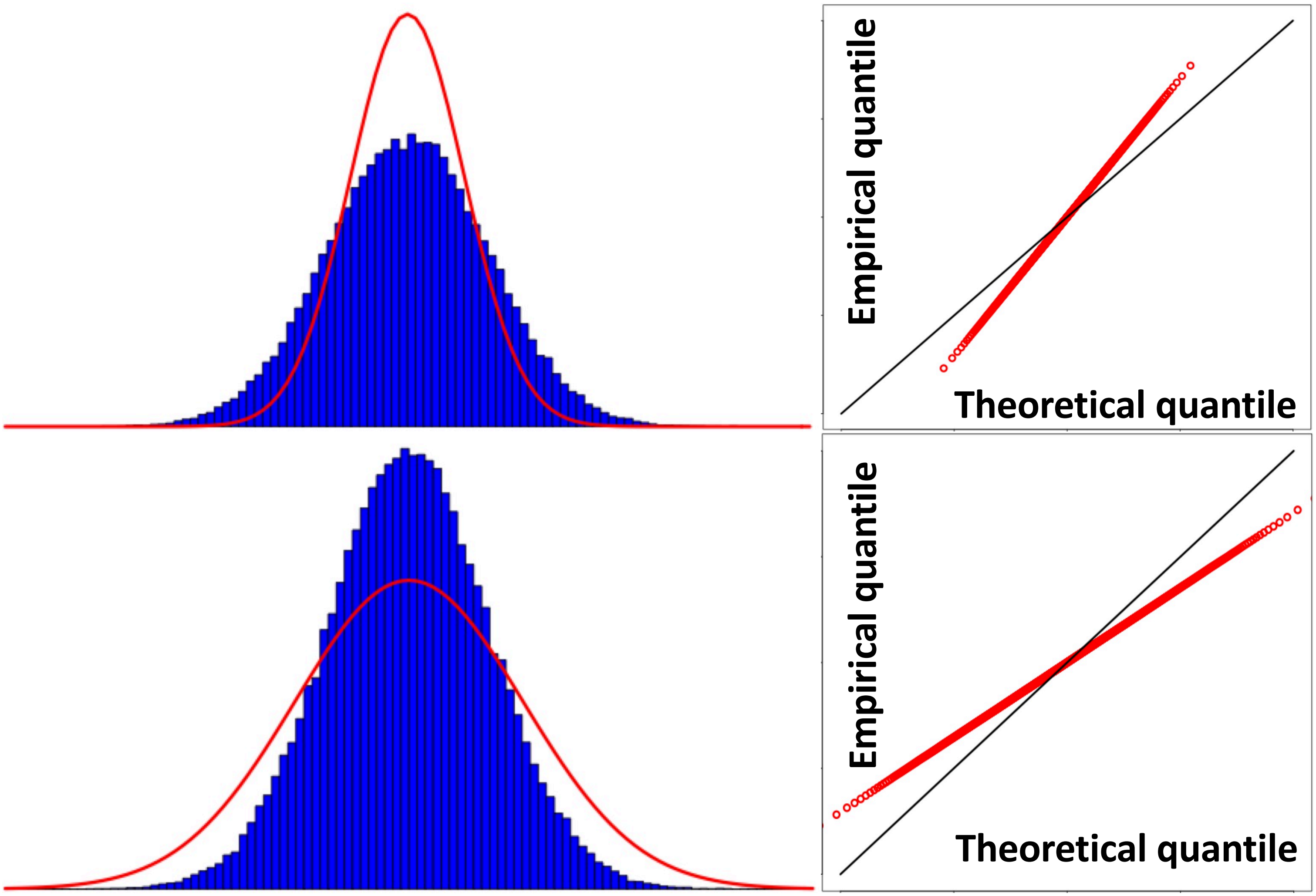


Q-Q plots



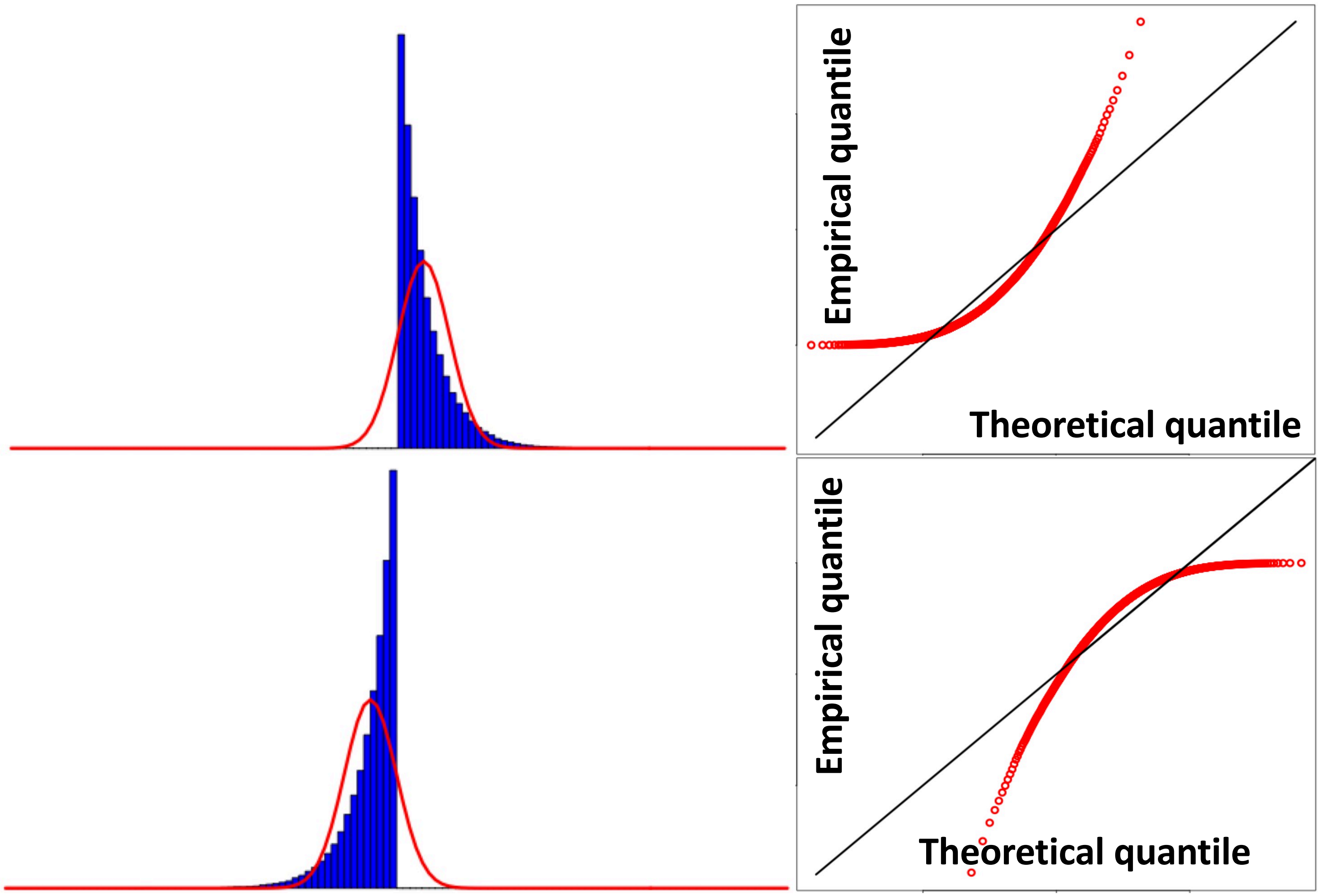
Incorrect mean introduces a constant offset in QQ plot

Q-Q plots



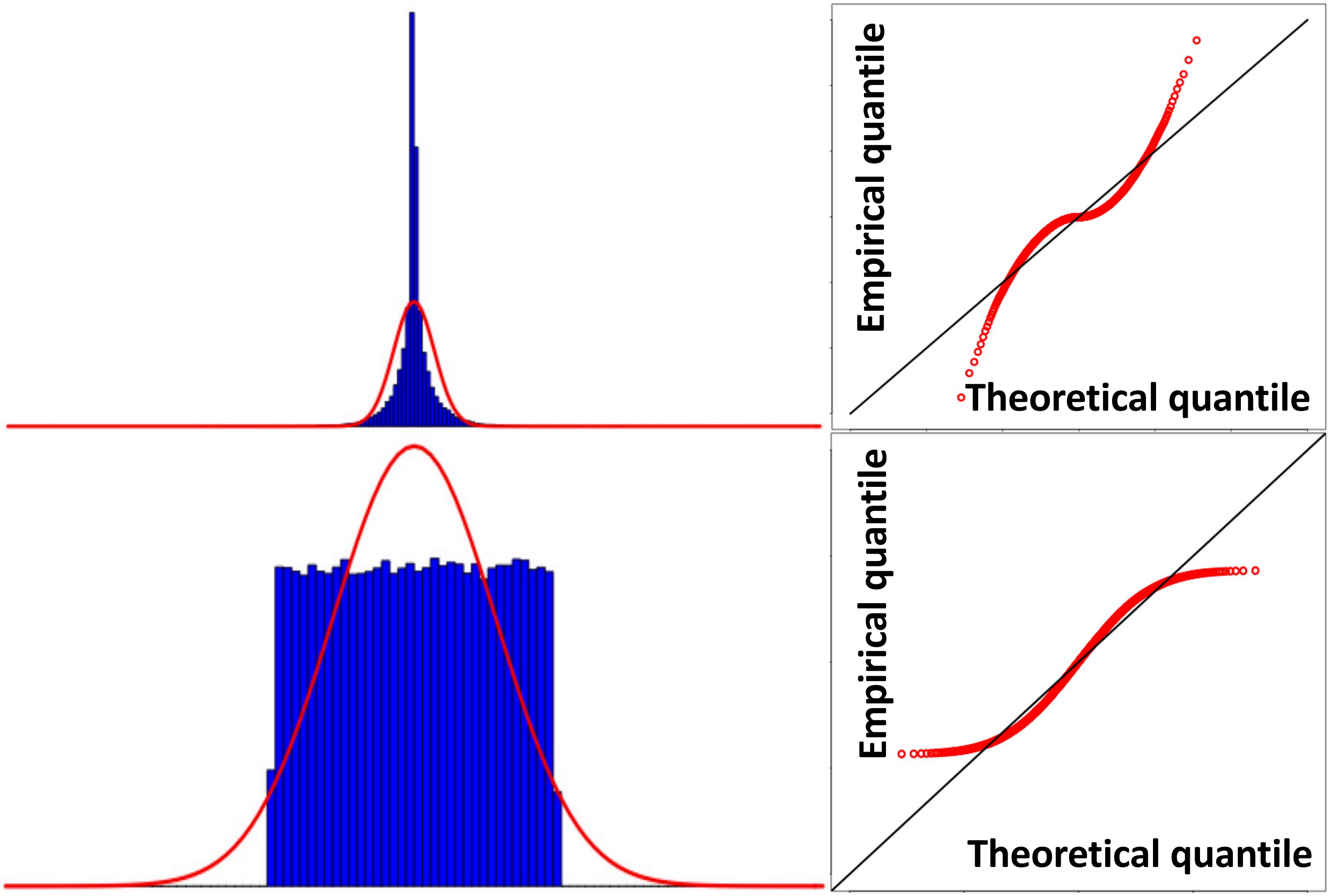
Incorrect variance introduces a linear slope in QQ plot

Q-Q plots



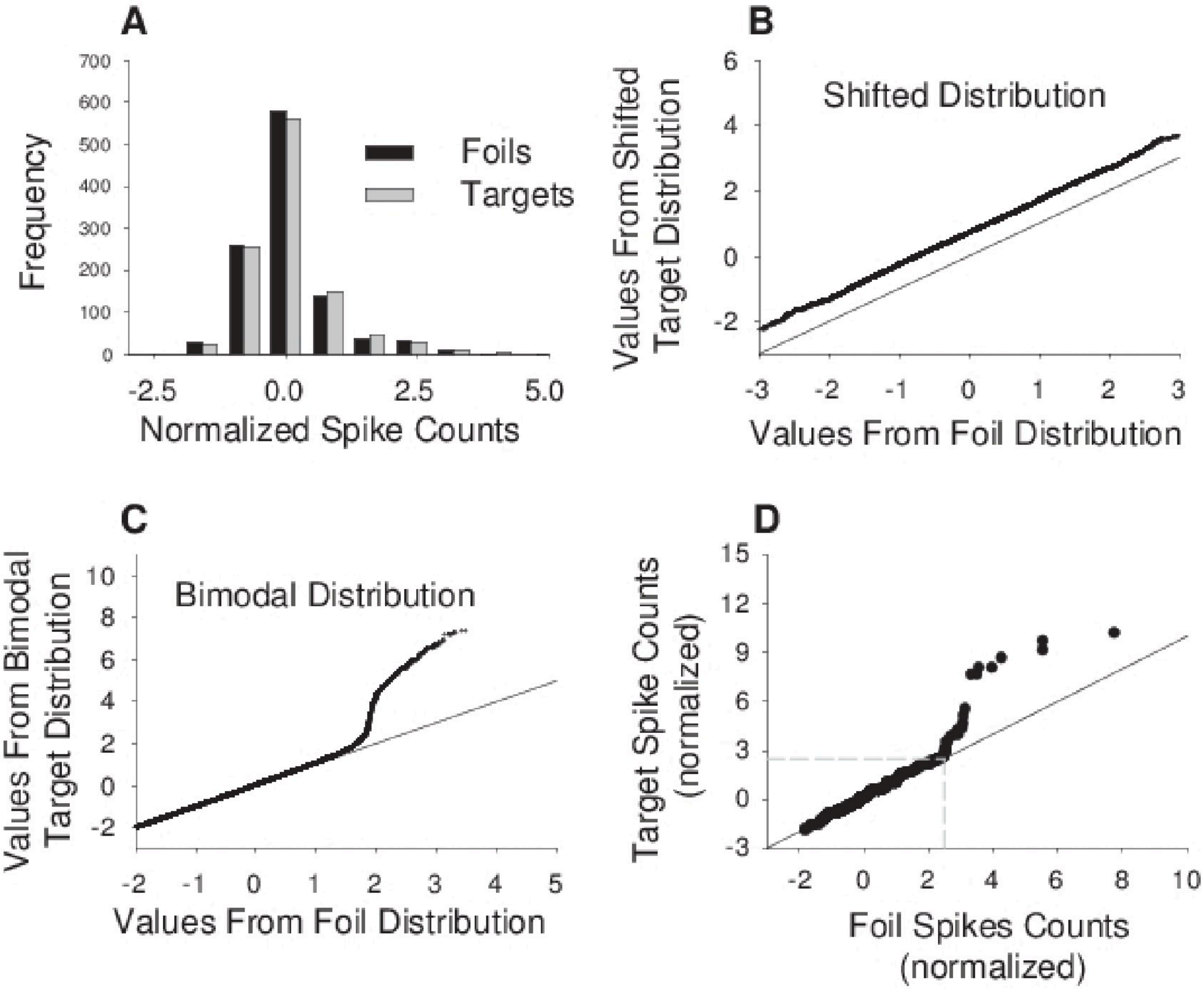
Incorrect skew adds a ‘quadratic’ deviation in QQ plot

Q-Q plots



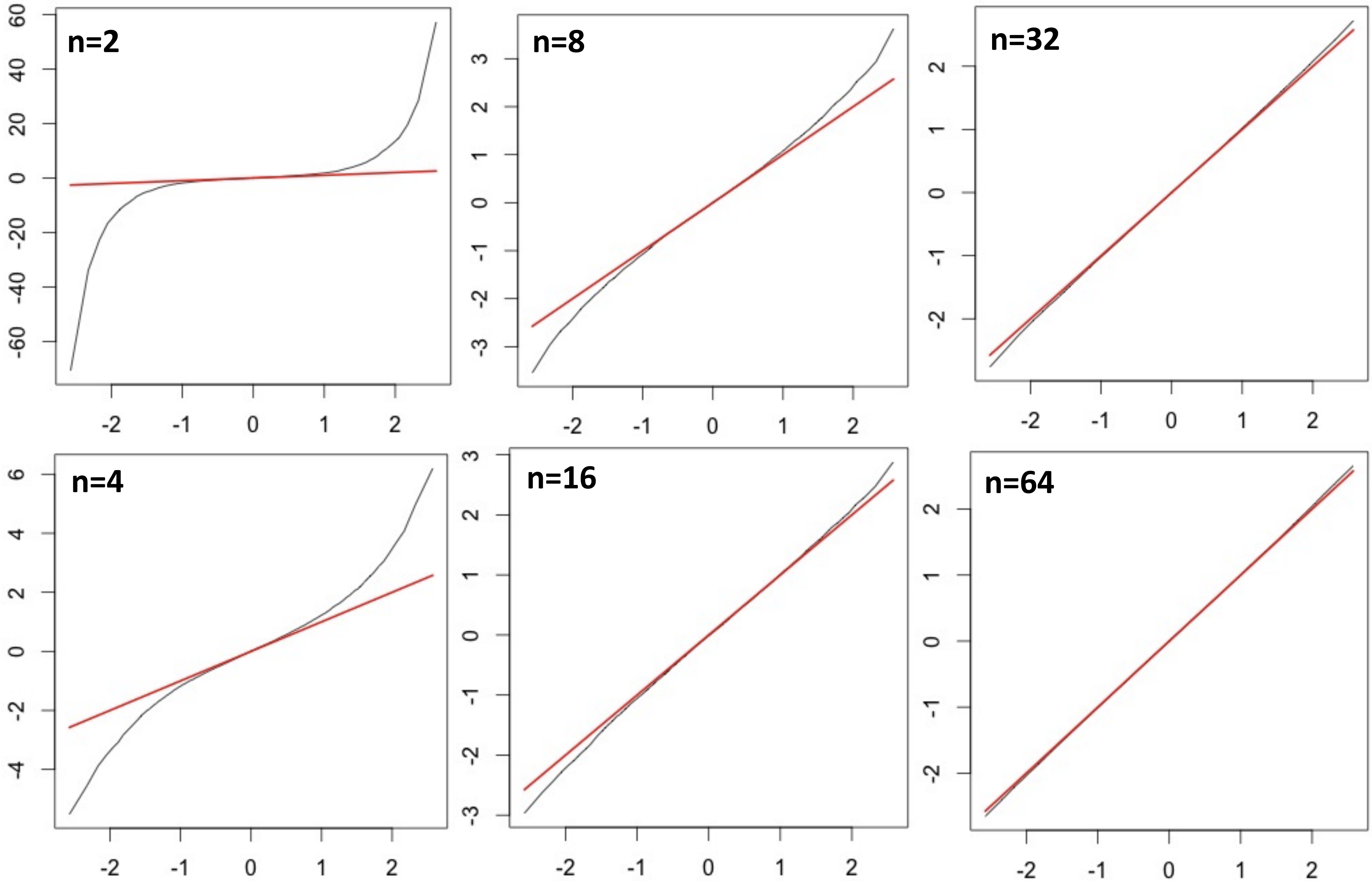
Incorrect kurtosis adds a ‘cubic’ deviation in QQ plot

Q-Q plots

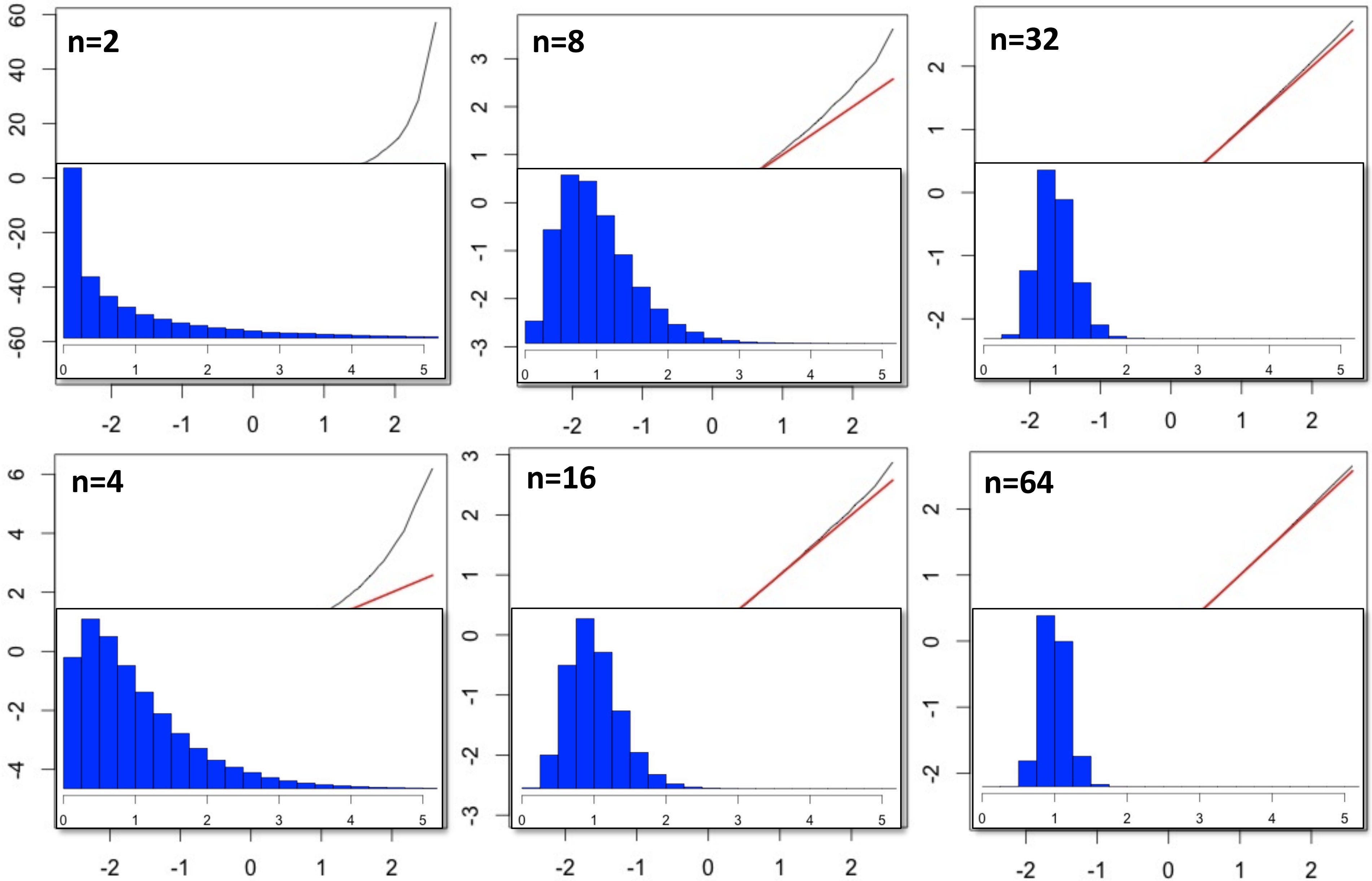


From John Wixted

QQ plots of **t-statistics** for samples with different ***n***s compared to std. normal distribution



QQ plots of **t-statistics** for samples with different ***n***s compared to std. normal distribution

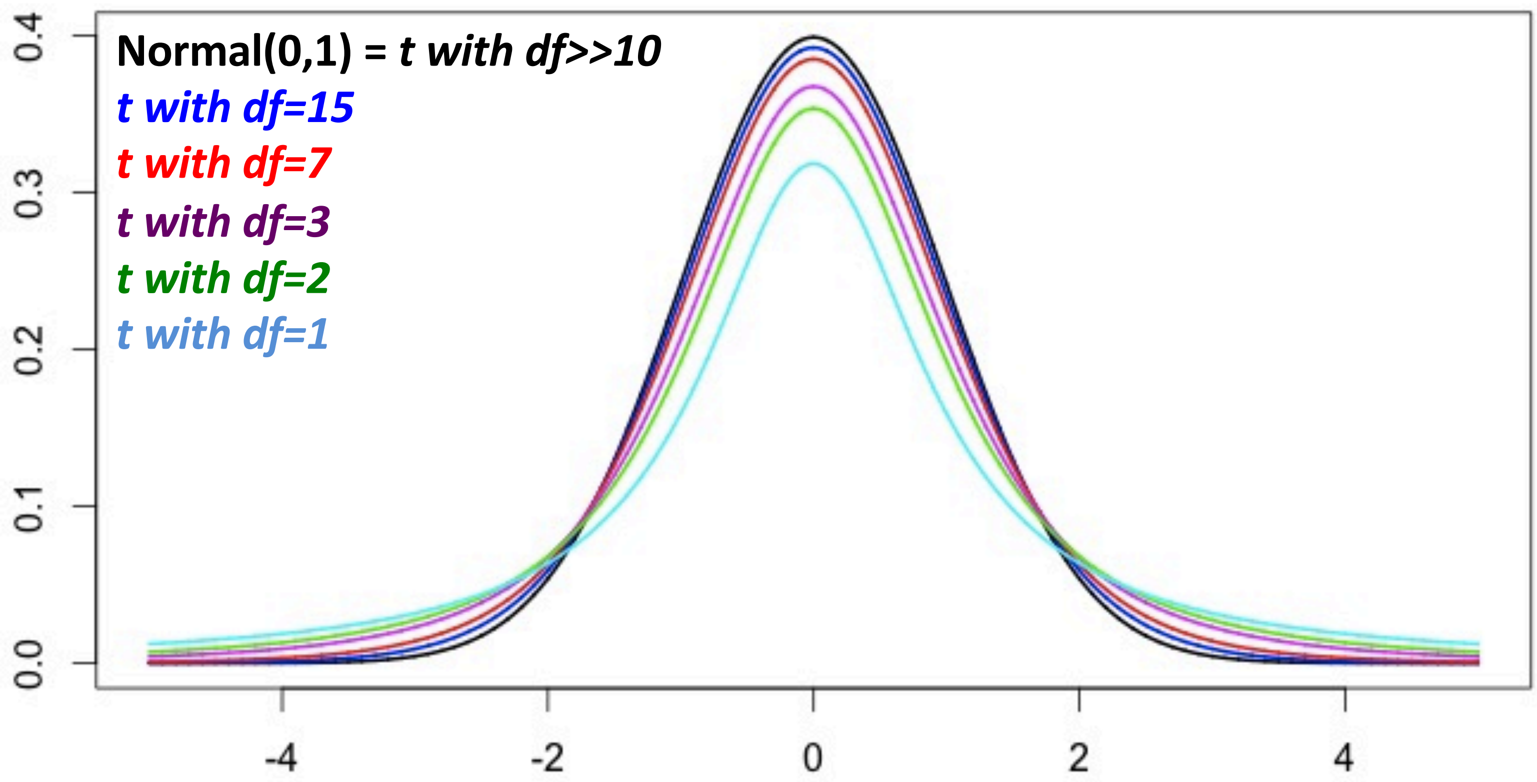


Degrees of freedom again?

$$\hat{\sigma} = s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2} \qquad t_{\overline{x}} = \sqrt{n}\left(\frac{\overline{x} - \mu_0}{s}\right)$$

df for the t-distribution: how many data points were free to vary when estimating the variance?

(if we estimated one mean, one data point was not free to vary, so df = n-1)



	Z-test	One-sample t-test	Paired t-test	2-sample eq. var. t-test	2-sample uneq var t-test
	We know pop. var. Want to test if mean differs from H0 mean.	We **do not know** pop. var. Want to test if mean differs from H0 mean.	We have 2 measures of the same thing, do they differ in means?	We want to know if two samples (assumed to have equal var) have different means	We want to know if two samples (not assumed to have equal var) have different means
Statistic	$z_{\bar{x}} = \left(\frac{\bar{x}-\mu_0}{\sigma_X}\right)\sqrt{n}$	$t_{\bar{x}} = \left(\frac{\bar{x}-\mu_0}{s_X}\right)\sqrt{n}$	$t_{\bar{D}} = \left(\frac{\bar{D}}{s_D}\right)\sqrt{n}$	$t_{\bar{x}-\bar{y}} = \frac{\bar{x}-\bar{y}}{s_P\sqrt{\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}}$	$t_{\bar{x}-\bar{y}} = \frac{(\bar{x}-\bar{y})}{\sqrt{\left(\frac{s_X^2}{n_x}+\frac{s_Y^2}{n_y}\right)}}$
Effect size	$\hat{d} = \left(\frac{\bar{x}-\mu_0}{\sigma_X}\right)$	$\hat{d} = \left(\frac{\bar{x}-\mu_0}{s_X}\right)$	$\hat{d} = \left(\frac{\bar{D}}{s_D}\right)$	$\hat{d} = \left(\frac{(\bar{x}-\bar{y})-\mu_0}{s_P}\right)$	Effect size here breaks the mold because of the diff. variances.
d.f.		$df = n-1$	$df = n-1$	$df = n_1+n_2-2$	$df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}}$
P-value 2-tailed	`2*pnorm(-abs(z))`	`2*pt(-abs(t),df)`			

Standard errors of the mean / difference

	Z-test	One-sample t-test	Paired t-test	2-sample eq. var. t-test	2-sample uneq var t-test
1-α% C.I.	$\bar{x} \pm z_{\alpha/2}\boxed{\sigma_X/\sqrt{n}}$	$\bar{x} \pm t_{\alpha/2}\boxed{s_X/\sqrt{n}}$	$\bar{D} \pm t_{\alpha/2}\boxed{s_D/\sqrt{n}}$	$(\bar{x}-\bar{y}) \pm t_{\alpha/2}\boxed{s_P * \sqrt{1/n_1+1/n_2}}$	$(\bar{x}-\bar{y}) \pm t_{\alpha/2}\boxed{\sqrt{\frac{s_X^2}{n_x}+\frac{s_Y^2}{n_y}}}$
	`z* = qnorm(a/2)`	`t* = qt(a/2,df)`			

Working from summary stats.

- Reported t-test: t(17)=2.5 t(df) = t.statistic
  - What's the two-tailed p-value?
- This is a paired-sample test
  - what's the sample size?
  - What's the (estimated) effect size?
- The mean of the difference was 5
  - What's the standard error of the difference?
  - What's the standard deviation?
  - What's a 95% confidence interval on the mean difference?

Working from summary stats.

- Reported t-test: t(28)=2.5 t(df) = t.statistic
  – What's the two-tailed p-value?
- This is an equal-variance, two-sample test, with matched sample sizes.
  – what's the sample size in each group?
  – What's the (estimated) effect size?
- The pooled sd was 10
  – What's the difference between means?
  – What's a 95% confidence interval on the difference in means?

T distribution

What is our confidence interval on….

- the mean math GRE score of psych students?
- the avg. improvement in math GRE scores from taking a Kaplan course?
- The difference in mean math GRE scores between psych and cog sci students?
- the difference in avg. improvement from taking a Kaplan course different vs doing some practice tests?

$$\bar{x} \pm t_{\alpha/2} \boxed{s_X / \sqrt{n}}$$

Because it's a one sample t-test, CI on the mean

$$\bar{D} \pm t_{\alpha/2} \boxed{s_D / \sqrt{n}}$$

CI on the mean difference – effectively 1-sample t-test.

$$(\bar{x} - \bar{y}) \pm t_{\alpha/2} \boxed{s_P * \sqrt{1/n_1 + 1/n_2}}$$

CI on difference between two means with eq var. so std. err. is different

$$(\bar{x} - \bar{y}) \pm t_{\alpha/2} \boxed{\sqrt{\frac{s_X^2}{n_x} + \frac{s_Y^2}{n_y}}}$$

CI on difference between two means with unequal var. so std. err. is different

D.F. for two sample unequal variance t-test

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \cdot \nu_1} + \frac{s_2^4}{N_2^2 \cdot \nu_2}} = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{s_1^4}{N_1^2 \cdot (N_1 - 1)} + \frac{s_2^4}{N_2^2 \cdot (N_2 - 1)}}$$

"Welch's t-test"

Intuition:

1) **When one standard error is way bigger than the other (either due to high variance, or low n), it's like doing a 1 sample t-test, because only the variance of that 1 sample will matter. So we want to have d.f. = n1-1**
2) **When the two standard errors are the same (or similar), it's like doing a 2-sample t-test, because both variances contribute equally. So we want d.f. = n1+n2-2**

This formula does that. t.test in r does this by default.