MATH 10

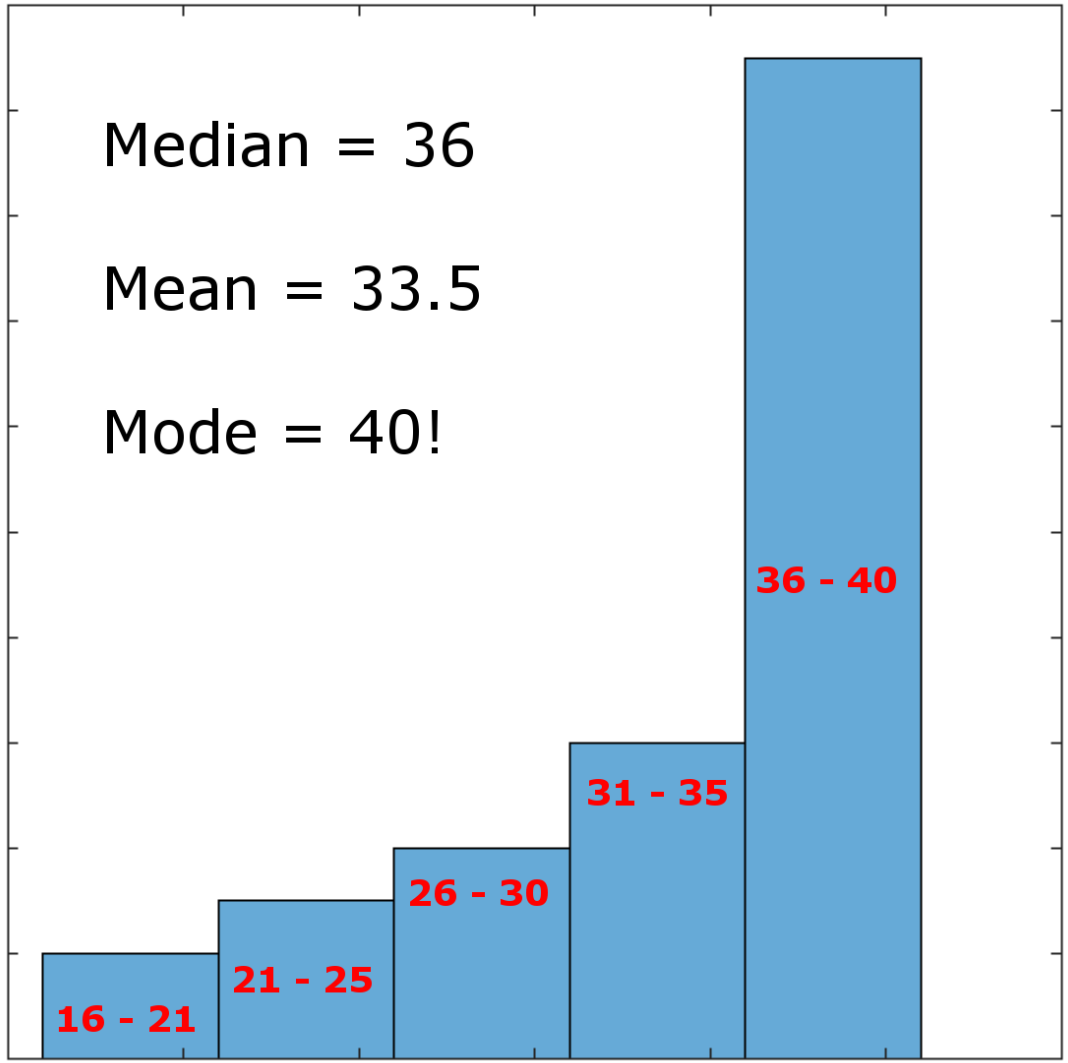
INTRODUCTORY STATISTICS

Tommy Khoo

Your friendly neighbourhood graduate student.

Midterm Exam

- The final exam for Math 10 is usually super tough.
- *Tentatively* looking at ~7 qns, 10 pts each.
- Cumulative. So you will see these types of questions again. 2.5 old - 4.5 new split maybe.



Week 6

- **Chapter 10 – Estimation** ← today's lecture

difference between means

- **Chapter 8 – Advanced Graphs** ← today's lecture

- **Chapter 11 – Logic of Hypothesis Testing**

FINALLY: significance testing, type I/II errors, one/two tailed tests etc.

Sampling Distributions and Confidence Intervals for Difference Between Means

- Will this be in the final exam and cost a lot of points? **ABSOLUTELY**
- Can be standalone questions, parts of other questions or two-in-one deal.

Sampling Distributions and Confidence Intervals for Difference Between Means

- Will this be in the final exam and cost a lot of points? **ABSOLUTELY**
- Can be standalone questions, parts of other questions or two-in-one deal.

- Recall the 4 major parts:

1) CLT and Normal distribution
2) t-distribution
3) Proportion
4) Difference between means → we will finally finish this

Chapter 9, Section 7 – Difference Between Means

- One way to compare two populations in statistics.

- Suppose you have two simple random samples with size n_1 and n_2.
- Samples from population 1 and 2 respectively.

- Calculate their sample means M_1 and M_2.
- The difference has a sampling distribution with mean

$\mu_{M_1-M_2} = \mu_1 - \mu_2$.

Chapter 9, Section 7 – Difference Between Means

- The difference has a sampling dist. with mean $\mu_{M_1-M_2} = \mu_1 - \mu_2$.
- And variance $\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_1}$.

- $\sigma^2_{M_i} = \frac{\sigma^2}{n_i}$, which is variance of the sampling dist. of M_i.
- Since the sample means are independent (as random variables), the variance sum law was used to derive the variance.

- $\sigma^2_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

Chapter 9, Section 7 – Difference Between Means

- The difference has a sampling dist. with mean $\mu_{M_1-M_2} = \mu_1 - \mu_2$.
- And variance $\sigma^2_{M_1-M_2} = \sigma^2_{M_1} + \sigma^2_{M_1} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$.

- Standard error $\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.

- This becomes much easier if the sample sizes and population variances are equal.
- $\sigma_{M_1-M_2} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}}$.

→ *exam trick: not factoring out the root 2*

Chapter 9, Section 7 – Difference Between Means

- We often do not have the variance σ^2.
- Like in the single population t-distribution case, we estimate with something call MSE.

- In the exam, you may encounter both cases. So watch out.
- The formula for MSE will be on the formula sheet.

- We will need a stack of assumptions to make calculating the MSE simple enough to be tested in Math 10.
- Tentative recipe: just replace all σ^2 with MSE.

Chapter 10, Section 8 – Confidence Interval for Difference Between Mean *FINALLY*

The textbook only deals with a particular case. The 4 assumptions for this case are:

1. Two populations have the same variance (homogeneity of variance).
2. Both populations are normally distributed.

Chapter 10, Section 8 – Confidence Interval for Difference Between Mean *FINALLY*

The textbook only deals with a particular case. The 4 assumptions for this case are:

1. Two populations have the same variance (homogeneity of variance).
2. Both populations are normally distributed.
3. Both simple random samples are completely independent.
4. Both simple random samples have the same size n.

We will stick to these assumptions every time we do differences between mean. (but know that the math actually works more generally)

Mean Square Error

1. Two populations have the same variance (homogeneity of variance).
2. Both populations are normally distributed.
3. Both simple random samples are completely independent.
4. Both simple random samples have the same size n.

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

Mean Square Error

1. Two populations have the same variance (homogeneity of variance).
2. Both populations are normally distributed.
3. Both simple random samples are completely independent.
4. Both simple random samples have the same size n.

$$MSE = \frac{s_1^2 + s_2^2}{2}$$

Then, (estimate of the) standard error is $s_{M_1 - M_2} = \sqrt{\frac{2MSE}{n}} = \sqrt{\frac{s_1^2 + s_2^2}{n}}$.

Chapter 9, Section 7 – Difference Between Means

Sample Exam Question (15 points)

You are doing research on the difference between the daily calories consumption of teenagers in the USA (population 1) vs those in Japan (population 2). You took independent simple random samples of size 100 from both populations.

Assumptions for this question:

1. Both populations are normally distributed, with the same **unknown** variance.
2. There are absolutely no problems with the two samples.

You are doing research on the difference between the daily calories consumption of teenagers in the USA (population 1) vs those in Japan (population 2). You took independent simple random samples of size 100 from both populations.

Assumptions for this question:

1. Both populations are normally distributed, with the same **unknown** variance.
2. There are absolutely no problems with the two samples.

A)

You want to know if teenagers in the USA consume more calories a day than teenagers in Japan. So, you constructed the variable: $\bar{X} = \bar{X}_1 - \bar{X}_2$, where $\bar{X}_i$ is the sample mean of population i. You also found these summary statistics:

Estimates of the sample variances are $s_1^2 = 185$ and $s_2^2 = 215$.

Population means are $\mu_1 = 2300$ and $\mu_2 = 2100$.

A)

You want to know if teenagers in the USA consume more calories a day than teenagers in Japan. So, you constructed the variable: $\bar{X} = \bar{X}_1 - \bar{X}_2$, where $\bar{X}_i$ is the sample mean of population i. You also found these summary statistics:

Estimates of the sample variances are $s_1^2 = 185$ and $s_2^2 = 215$.

Population means are $\mu_1 = 2300$ and $\mu_2 = 2100$.

What is the sampling distribution of $\bar{X}$? State the mean and standard error. *(6 pts)*

1 pt for sampling distribution, 2 pts for mean, 3 pts for standard error

Answer: variance not given, population normal. So, t distribution with degrees of freedom df = (100-1) + (100-1) = 198. mean = 2300 – 2100 = 200. SE = sqrt(2 MSE / n) = sqrt(4) = 2.

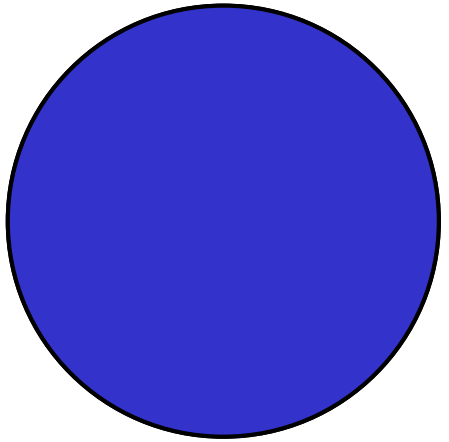
Break time!! \o/

- No in-class exercise today. ☹

- Circle is a timer that becomes blue. O_o →

(please ignore if it glitches)

12 minutes



Chapter 10, Section 11 – Confidence Intervals for Difference between Means *FINALLY*

- Formula:

$$[\,(\bar{X}_1 - \bar{X}_2) - t \cdot s_{M_1 - M_2}\,, (\bar{X}_1 - \bar{X}_2) + t \cdot s_{M_1 - M_2}]$$

- I will verbally explain how we obtained this formula.

A)

You want to know if teenagers in the USA consume more calories a day than teenagers in Japan. So, you constructed the variable: $\bar{X} = \bar{X}_1 - \bar{X}_2$, where $\bar{X}_i$ is the sample mean of population i. You also found these summary statistics:

Estimates of the sample variances are $s_1^2 = 185$ and $s_2^2 = 215$.

Population means are $\mu_1 = 2300$ and $\mu_2 = 2100$.

What is the sampling distribution of $\bar{X}$? State the mean and standard error. *(6 pts)*

1 pt for sampling distribution, 2 pts for mean, 3 pts for standard error

B) *Tricky* (°ل͜°)

Construct an interval, symmetric around the mean, using your answer in part A), so that the probability of getting a **new** difference in sample means within this interval is 99%. *(4 pts)*

B) *Tricky* (°ʖ°)

Construct an interval, *symmetric around the mean*, using your answer in part A), so that the probability of getting a **new** difference in sample means within this interval is 99%. *(4 pts)*

Use the formula $[\ \mu_1 - \mu_2\ \pm t\ SE\ ]$.

C)

Now you are given that $\bar{X}_1 = 2100$ and $\bar{X}_2 = 2000$.

Construct a 99% confidence interval for the difference between the means of population 1 and 2. *(5 pts)*

Use the formula $[\ \mu_1 - \mu_2\ \pm t\ SE\ ]$.

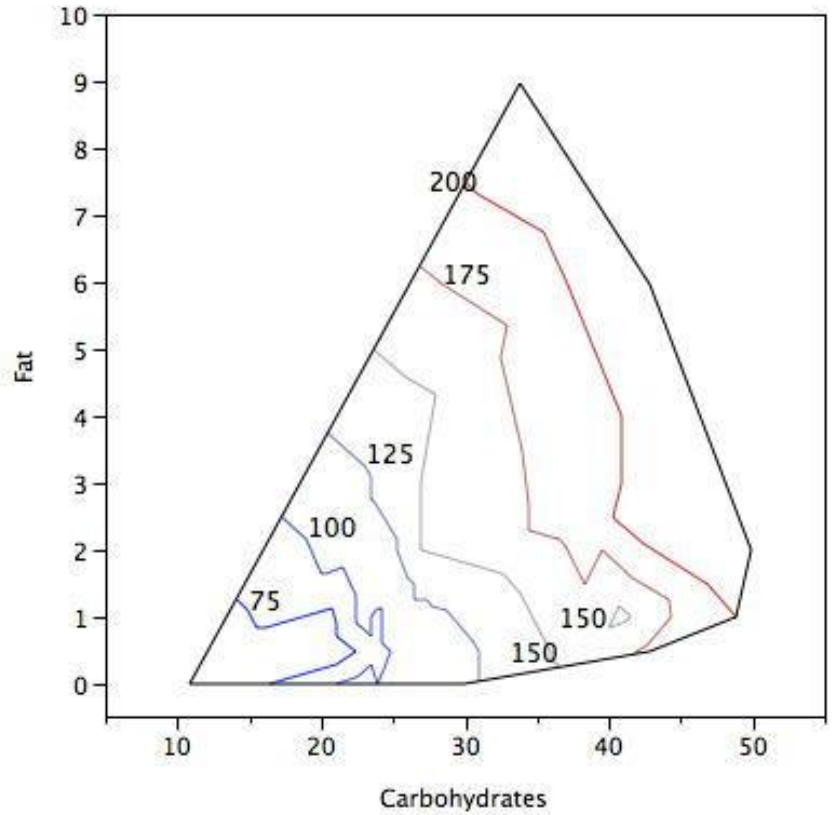
Let’s end the class with something light and easy.

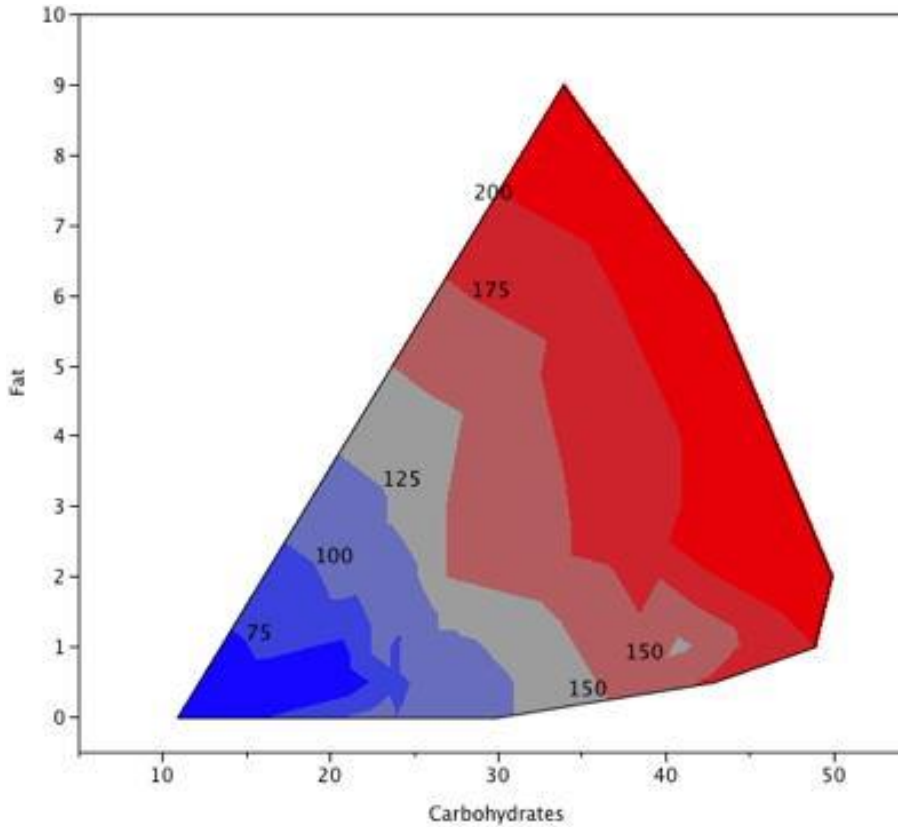
Chapter 8 – Advanced Graphs.

You might be asked to read Q-Q plots in the exam to come up with the obvious conclusion.

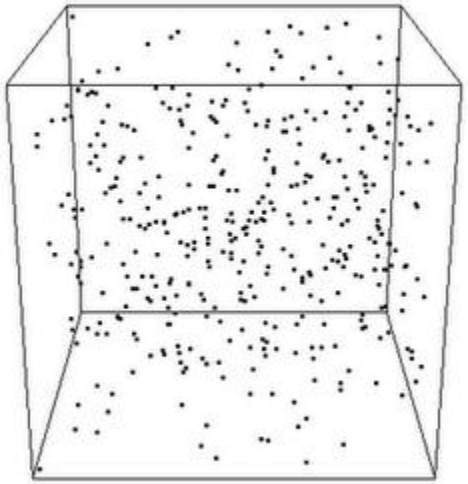
But not draw them.

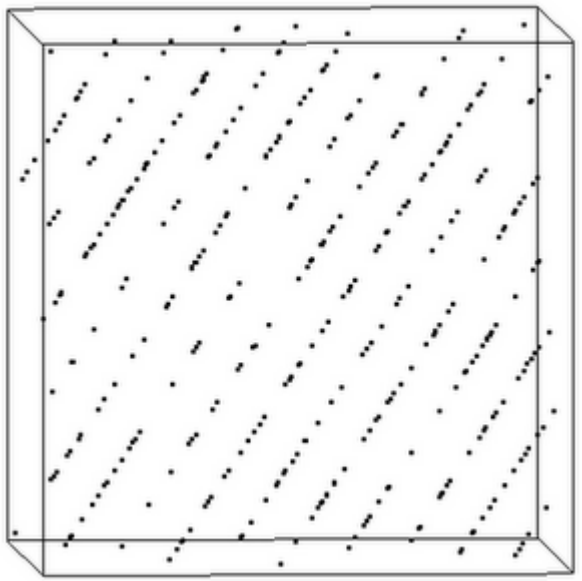
Chapter 8, Section 3 – Contour Plots

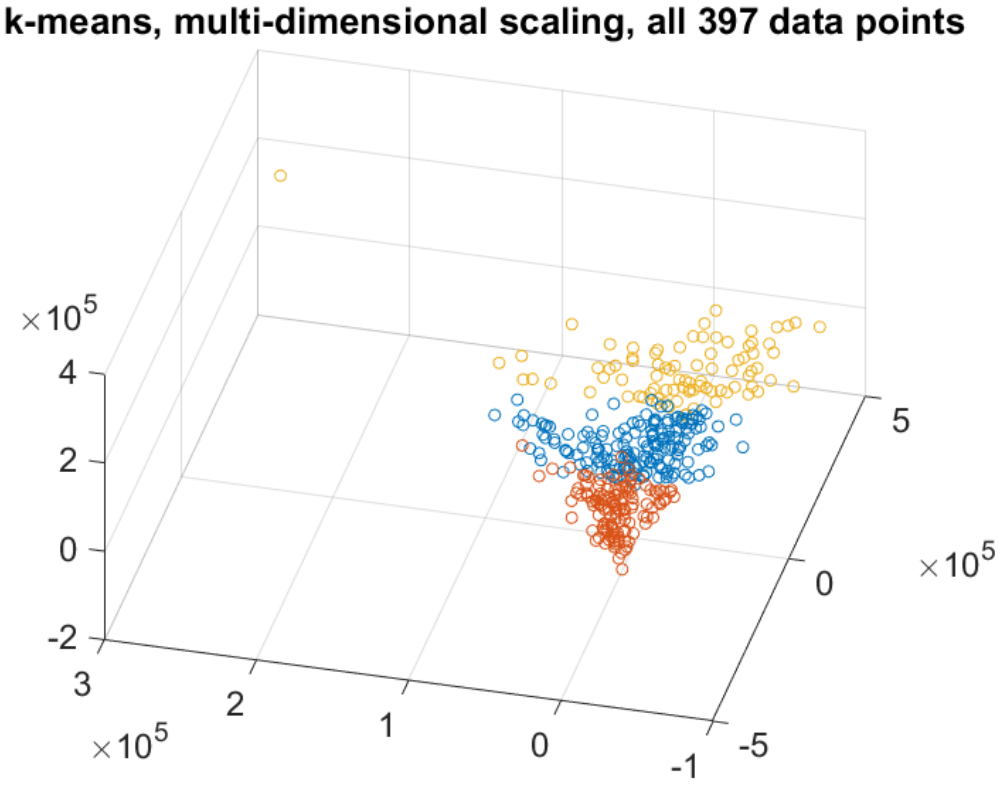




Chapter 8, Section 3 – 3D Plots

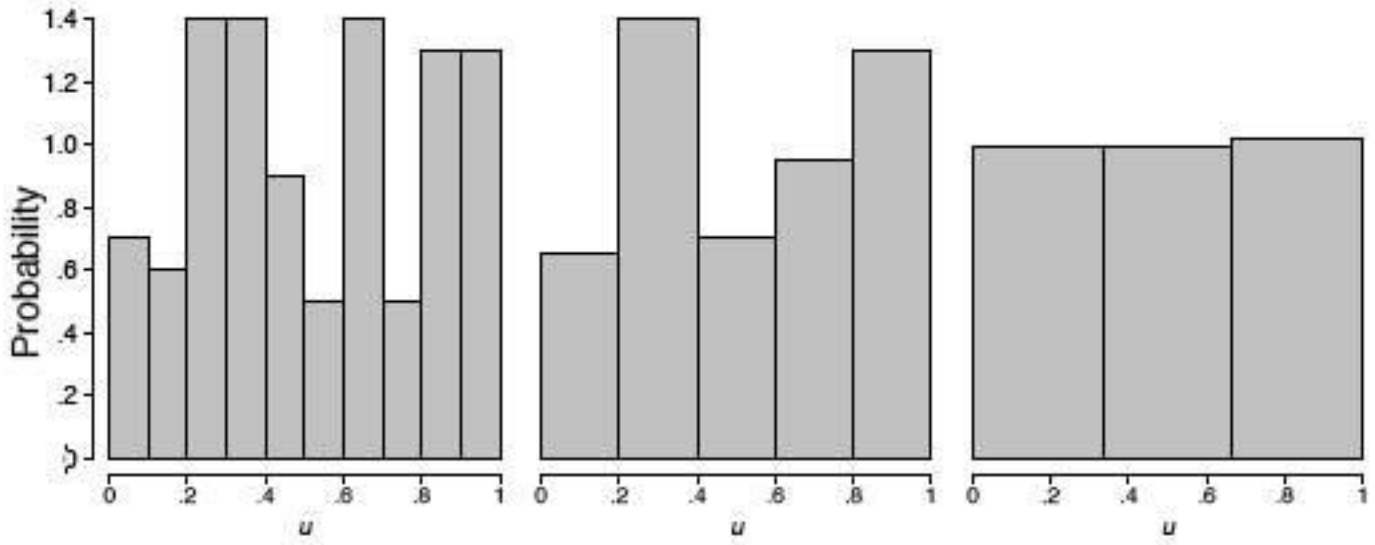






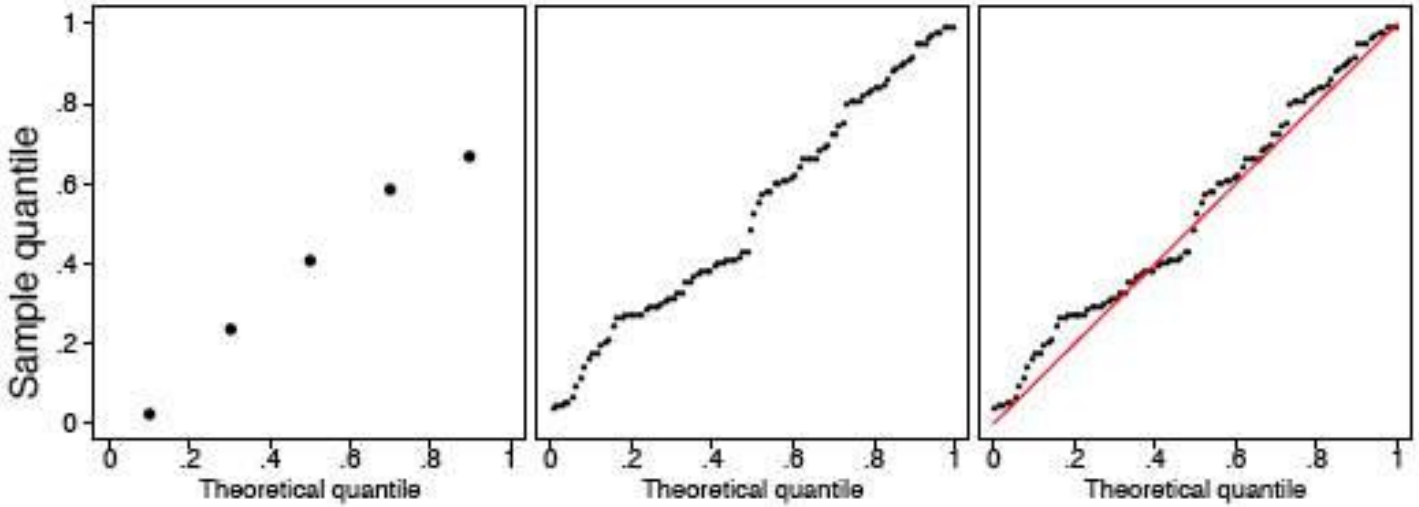
Chapter 8, Section 2 – Q-Q plots

- Very useful in applications! Basic idea: compare the quantiles of a theoretical distribution (normal, uniform etc) with the quantiles in your sample/data.
- **Note**: this section has a lot of technical details that are not expected of you in this course. What we do expect of you is the ability to read a Q-Q plot.
- The problem with just using histograms: it depends on the choice of bins/classes.



Chapter 8, Section 2

Q-Q plots



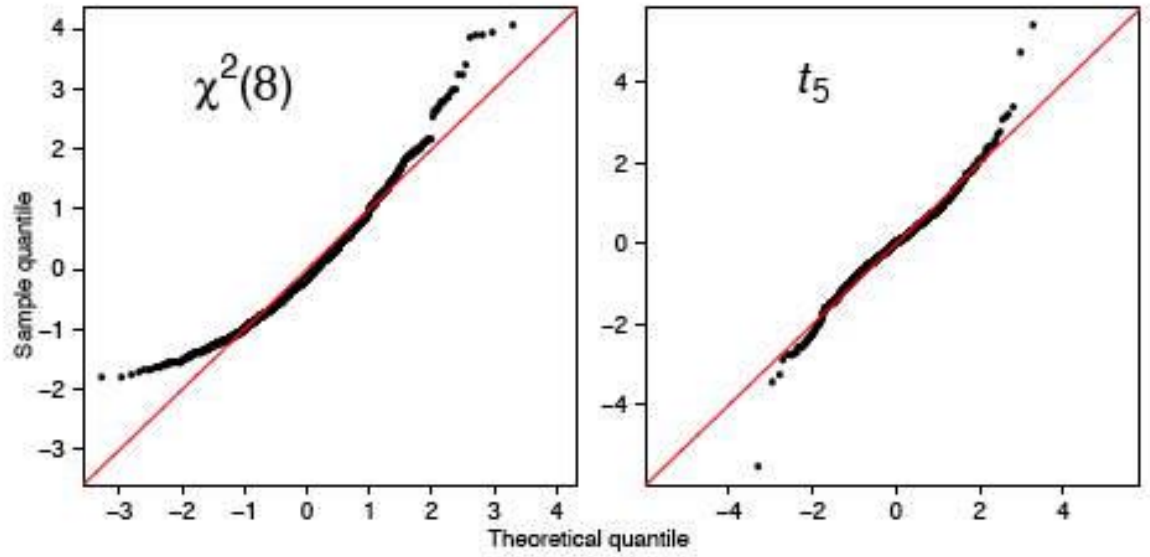
- Comparing theoretical and sample quantiles.
- Two cases for our course: uniform and normal data.

- qth quantile of n data points = a number such that (q times n) of the data is less than the number.
- E.g. 0.5^{th} quantile = median.

- Uniform distribution: order the n data points, then plot [(i-0.5)/n, i-th data].

Chapter 8, Section

Q-Q plots



- Convert normally distributed data to standard normal.
- Standard normal: order the n data points, then plot

[z-value such that P(Z <= z-value) = $(i - 0.5)/n$, z-value of the i-th data].

→ not in math 10 but what you're doing is actually marking the [0,1] interval at every $(i - 0.5)/n$, inverting the standard normal cumulative distribution function (CDF), which gives you the z-value the i-th data theoretically should be.

Q-Q Plot Exam Questions

- Will be a very tiny part of the exam, or not show up at all. ("fringe" topic)
- Will be very obvious.
- E.g. look at the Q-Q plot on the left. Using the statistics from this course (and nothing else), would you say the normal distribution is a good approximation for our data? Explain. (*2 pts*)

Just follow this recipe:

1. Are the dots close to the red line?
2. Yes, very close => yes, good approximation.
3. No, quite a few are far away! => no, bad approximation.